A Codebook-Based Modeling Approach for Bayesian STSA Speech Enhancement

Golnaz Ghodoosipour

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

May 2014

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Engineering.

Abstract

Speech enhancement algorithms are a fundamental component of digital speech and audio processing systems and currently find applications in a wide variety of consumer products for storage, transmission and playback of voice, including: cell phones, video cameras, PDAs voice recorders, teleconference speaker phones and hands-free car phones. Over the last few decades, the problem of speech enhancement has been vastly studied in the technical literature because of the increasing demand for removing a certain amount of background noise from the desired speech signal.

Different approaches have been proposed for the enhancement of speech contaminated by various types of noise. The common goal is to remove as much noise as possible without introducing distortion to the processed speech. Among the different categories of speech enhancement methods, frequency-domain approaches are usually favored in applications due to their lower complexity, ease of implementation on a real-time digital signal processor and resemblance to the natural processing taking place in the human auditory system. Within the family of frequency-domain approaches, Bayesian estimators of the short-time spectral amplitude (STSA) offer the best overall performance in terms of noise reduction and speech distortion. While the STSA methods have been successful under stationary noise conditions, the problem of speech enhancement in a nonstationary noise environment is still an open issue for research.

The main goal of this thesis is to develop a Bayesian STSA estimator with the purpose of single-channel speech enhancement in the presence of moderate levels of nonstationary noise. In this regard, we use a Bayesian minimum mean squared error (MMSE) approach for the joint estimation of the short-term predictor parameters of speech and noise, from the noisy speech observation. This approach is based on a recent work by Srinivasan et al. where trained codebooks of speech and noise linear predictive (LP) coefficients are used to model the *a priori* information required by the Bayesian MMSE estimation. Afterwards, the estimated power spectra are passed to the Wβ-SA Bayesian STSA speech enhancement method, where they are used to calculate the enhancement gain in the frequency domain. Finally, these gains are applied to the noisy speech short-term Fourier transforms. which are then converted back to the time-domain to obtain the desired estimate of the clean speech. When compared to an existing benchmark approach from the literature, the proposed speech enhancement approach developed in this thesis gives rise to a notable improvement in the quality of the processed noisy speech.

Sommaire

Le rehaussement numérique de la parole est une composante fondamentale des systèmes de traitement audio et trouve actuellement des applications dans une vaste gamme de produits de consommation pour l'entreposage, la transmission et la reproduction de la voix, y compris : les téléphones cellulaires, caméras vidéo, enregistreurs vocaux PDA (assistants numériques), systèmes de téléconférence et téléphones mains-libres d´automobile. Au cours des dernières décennies, le problème du rehaussement de la parole a été considérablement étudié dans la littérature technique en raison de la demande croissante pour la réduction lu niveau de bruits de fond à partir du signal vocal désiré dans ces applications.

Différentes approches ont été proposées pour le rehaussement de parole contaminée par différents types de bruit. L'objectif commun est de supprimer autant de bruit que possible sans introduire de distorsion au signal parole. Parmi les différentes catégories de méthodes proposées pour l'amélioration de la parole, les approches dans le domaine fréquentiel sont généralement favorisées en raison de leur complexité inférieure, la facilité de mise en œuvre sur un processeur numérique en temps réel et la ressemblance avec le traitement naturel ayant lieu au sein du système auditif humain. Dans la famille des approches fréquencielles, les estimateurs bayésiens de l'amplitude spectrale à courte durée (STSA) offrent la meilleure performance globale en termes de la réduction du bruit et la distorsion de la parole. Alors que les méthodes STSA ont réussi dans les conditions de bruit stationnaire, le probléme de l'amélioration de la parole dans un environnement de bruit non-stationnaire est encore une question d'intérêt courant pour la recherche.

Le principal objectif de cette thèse est de développer une estimation bayésien amélior´ee des paramètres STSA dans le but de rehausser la qualité d'un signal parole (canal unique) en présence de niveaux modérés de bruits non-stationnaires. À cet égard, nous utilisons une formulation bayesienne basé sur la minimisation de l'erreur quadratique moyennede des paramètres à prédictifs à court terme de la parole et du bruit, partir de l'observation de la parole bruitée. Cette approche est fondée sur un travail récent par Srinivasam *et al.* dans lequel des livres de codes sont utilisés pour la représentation des coefficients de prédiction liné (LP) et gains d'excitition de la parole et du bruit. Ces livres de codes sont à leur tour utilisés afin de réaliser l'estimation MMSE des spectres de puissance qui sont requis lors de l'application de la méthode de rehaussement STSA. Dans cette thèse, les spectres de puissance estimés par l'approache MMSE sont utilisés au sein de la méthode Wβ-SA, où ils servent à calculer le gain de rehaussement qui sera appliqué au signal btuité dans le domaine de fréquence. En comparaison avec une méthod existante, la nouvelle méthode de rehaussement de la parole proposée dans cette thèse donne lieu à des améliorations importantes de la qualité du signal.

Acknowledgment

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Benoit Champagne, for his continuous support. With no doubt, this thesis would not have been possible without his inspiration, constructive advice and lots of helpful ideas.

I want to dedicate this thesis to my father, may his soul rest in peace, who passed away during the time I was working on my thesis. Thank you for giving me support and love in all stages of my education, and help me accomplish all that I have.

I am grateful for the financial support provided by Prof. Champagne via his research grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and Microsemi Canada Ltee without which this thesis would not have been possible.

I also acknowledge the help provided by Prof. Eric Plourde in the city of Sherbrooke for providing useful guidance and practical ideas.

I must express my profound gratitude to my mother for providing me with unfailing support and continuous encouragement throughout my years of study, my brother, Farzad, who has been a source of encouragement and inspiration throughout my graduate studies in Canada, my sister, Farnaz, for her incomparable love and kindness beside me in Canada and my other sister, Behnaz, who has always been as my best friend.

This journey would have been much more difficult without my best friends, Samira, Shohreh, Ghazaleh, Bahareh, Golnaz, Niloufar, Katayoun, Mahdi, Ahmad, Hessam, Dena and Mohammad. I am also grateful to my fellow colleagues in the Telecommunications and Signal Processing laboratory.

Contents

List of Figures

List of Tables

List of Acronyms

SNR	Signal to Noise Ratio
PSD	Power Spectral Density
VAD	Voice Activity Detector
MCRA	Minima Controlled Recursive Averaging
IMCRA	Improved Minima Controlled Recursive Averaging
SM	Single Microphone
MA	Microphone Array
STFT	Short-Time Fourier Transform
KLT	Karhunen-Loeve Transform
DCT	Discrete Cosine Transform
FFT	Fast Fourier Transform
MMSE	Minimum Mean Square Error
STSA	Short-Time Spectral Amplitude
WE	Weighted Euclidean
PDF	Probability Density Function
ML	Maximum Likelihood
STP	Short Time Predictive
HMM	Hidden Markow Model
AR	Auto-Regressive
LP	Linear Predictive
DFT	Discrete Fourier Transform
VQ	Vector Quantization
GLA	Generalized Lloyd Algorithm
STP	Short Time Predictor

LLF	Log Likelihood Function
PESQ	perceptual evaluation of speech quality

Chapter 1

Introduction

This chapter provides a general introduction to the thesis, which aims at developing and studying signal processing algorithms for the problem of speech enhancement in nonstationary environments. A high level overview of speech enhancement and its applications is given in Section 1.1, while a literature review of various speech enhancement methods and algorithms is presented in Section 1.2. The research objectives and the contributions of the thesis are discussed in Section 1.3, and finally, an outline of the upcoming chapters is presented in Section 1.4.

1.1 Speech Enhancement in Modern Communications Systems

1.1.1 What is speech enhancement?

Speech communications refer to the transmission of information from a speaker to a listener in the form of intelligible acoustic signals produced by the speaker vocal tract [1]. While it is the most effective and natural way for human beings to communicate, in today's busy world where noise is almost always present and silence rarely happens, the speech signal at the input of a communication system is usually degraded by various types of acoustic noises. The transmission of this signal can be through the air, i.e. directly from the speaker to the listener, or via electronic means including optical fibers, copper wires or radio waves [2]. The acoustic noise contaminates the speech and depending on its level, impairs the ability to communicate naturally or even reliably.

In all the applications of speech communications and speech processing, additive noise is present and degrades the quality and performance of the underlying system. Examples of such applications include sound recording, cell phones, hands-free communications, teleconferencing, hearing aids, and human-machine interfaces such as an automatic speech recognition system [3]. The noise corrupting the signal affects human-to-human as well as human-to-machine communications directly. The presence of acoustic noise poses a major problem to the system design, since it may cause significant changes in the speech signal characteristics. On the listener (i.e., receiver) side, the noise adds to the received signal and changes its spectral and statistical properties. However, changes may even occur on the speaker (i.e., transmitter) side where the talker tends to change his style in response to a high level of background noise [3].

Generally, regardless of exactly how the noise changes the speech characteristics, low to moderate level of noise corrupting a speech signal will lower its perceptual quality for the listener or the processing device, while high level of noise may degrade its intelligibility or render the processing ineffective. Therefore, the process of cleaning up the noisy speech signal at either the transmitting or the receiving end of the communication chain is highly desirable, and sometimes absolutely necessary. The cleaning process, which is often referred to as either *speech enhancement* or *noise reduction*, has become a crucial area of study in the field of speech processing [4].

Over the last few decades, the problem of speech enhancement has been studied vastly in the technical literature. With the emergence of cheap and reliable digital signal processing hardware, many powerful approaches and methods have been developed in order to remove a certain amount or types of noise from a corrupted speech signal. In general, these methods aim to achieve three main goals. The first one is to improve the perceptual quality of the noise-corrupted speech, as measured by various objective performance metrics such as the signal-to-noise ratio (SNR). Secondly, they aim to improve the speech intelligibility which is mainly a measure of how comprehensible is the speech. The third objective is to improve the performance of subsequent processing functions, such as speech coding, echo cancellation and speech recognition [3].

Most, if not all, speech enhancement approaches reported in the literature attempt to reduce the noise to an acceptable level while preserving the naturalness and intelligibility of the processed speech. However, there is always a trade off between these two conflicting objectives and it is often necessary to sacrifice one at the expense of the other [1]. An overview of the existing speech enhancement methods that are relevant to this project will be presented in Section 1.2.2 and 1.2.3.

1.1.2 What makes it difficult?

Today's speech communication systems are used in adverse acoustic environments, where various types of noise, interference and other undesirable effects may impair the quality and naturalness of the desired speech. The different physical mechanisms responsible for degrading the quality of a desired speech signal can be classified into four different categories [3]: additive noise, echo, reverberation and interference. Additive noise usually refers to natural sounds from unwanted acoustic sources (e.g. fan noise, traffic, etc.) or artificial sounds such as comfort noise in speech coder. These noise sources combine additively to the desired speech and change the details of its waveform. Echo is the phenomenon in which a delayed and distorted version of an original sound or electrical signal is reflected back to the source. In hands free telephony instance, echo usually occurs because of the coupling between loudspeakers and microphones [5]. In the case of echo, these reflections can be resolved or identified by the human auditory system. Reverberation is conceptually similar in that it is produced by reflection of a sound wave on walls and other objects, but in this case the reflected sound waves are so dense and closely spaced in time that they cannot be resolved by the auditory system. They are associated to the exponentially decaying tail of the acoustic impulse response between the source (speaker) and the destination (listener or microphone), which in turn is a consequence of the multiple reflections and absorption of the acoustic waves by the surrounding objects and surfaces. Finally, interference happens when multiple competing speech sources are simultaneously active, such as in teleconferencing or telecollaboration applications [3]. In this thesis, the main focus is on the enhancement of speech contaminated by additive noise and especially background acoustic noise.

One of the main challenge in speech enhancement is that the nature and characteristics of the additive noise change from one application to another. The problem is even more difficult when the statistical characteristics of the noise degrading the speech change over time in a given application [3]. Indeed, when the additive noise exhibits such as nonstationary behavior, the speech processing system must be able to track the frequent changes in the noise, and it becomes difficult to estimate its statistics which are needed as part of the enhancement process.

Another important and challenging issue is the ever present trade-off between noise reduction and speech distortion. Indeed it is invariably found that reducing the additive noise present in a speech signal introduces undesirable changes (distortion) to the latter. Modern approaches of speech enhancement often include design parameters which can be adjusted to control this trade-off. This means that the speech enhancement system should work in such a way as to achieve balance between reducing the amount of noise and degrading the speech quality .

Overall, the various methods of speech enhancement developed over the years, have reached an acceptable level of performance under a limited range of operating conditions, especially for a low level of stationary or non-stationary noise. However the enhancement of speech corrupted by high level levels of noise, especially non-stationary, remains an open problem for research. Below, we provide an overview of existed methods of speech enhancement indicating their advantages and their drawbacks. A more detailed description of selected speech enhancement and related noise estimation algorithms which are more closely to this work are given in Chapter 2.

1.2 Literature Review

Speech enhancement techniques have been amply studied and a wide range of algorithms operating under different conditions have been proposed. In all these approaches, the enhancement made to the noisy speech depends on the statistical properties of the desired speech and of the corrupting noise, which must be estimated as part of the enhancement process. A crucial component of a functional speech enhancement system, therefore is the estimation of the background noise statistics. Consequently, many algorithms have been developed for this purpose. An overview of which is therefore given in Section 1.2.1. This is followed by a review of speech enhancement methods in Sections 1.2.2 and 1.2.3, where in the latter section, the focus is on methods that employ statistical learning approaches.

1.2.1 Estimation of the noise statistics

The requirement for accurate estimates of the noise statistics is a common feature in most speech enhancement systems. Indeed the noise statistics are needed as part of the algorithm employed to clean the noisy speech. An example of this is in the calculation of optimum gains based on a probabilistic noise model for the filtering of the noisy speech. Typically, these gains require the knowledge of the short-time power spectral density (PSD) of the noise. The main problem here is that the noise statistics must be estimated from the noisy speech data, i.e. in the presence of the desired speech.

The most common noise estimation algorithms can be classified into two main families, namely hard-decision and soft-decision methods. In the first family, the noise statistics are tracked only during silence or noise-only periods of the noisy speech data, i.e. when the speech is inactive. This requires the use of a so-called “voice activity detector” (VAD) which apply some hypothesis tests based on certain energy measures [6], [7], [8]. However, estimating the noise statistics only during speech silence is not adequate in the case of a non-stationary noise environment, where the noise power spectral density (PSD) may change notably during a period of speech activity. Therefore, there is a need for noise estimation methods in which the noise PSD estimates are updated more frequently.

In the second family, referred to as soft-decision methods, the noise statistics are tracked even during speech activity. In recent years, several noise estimation algorithms have been proposed that fit into this category. These can be further divided into different subsets depending on their fundamental principle of operation. In a first, and possibly most important subset, the estimates of the noise statistics are obtained through a minimum controlled process, as exemplified by [9], [10], [11]. A short description of these algorithms is given below.

In [9], Martin proposed an original method for estimating the noise PSD, which is based on tracking the minimum of the noisy speech short-term PSD over a finite temporal window. This comes from the observation that the power level of a noisy speech signal frequently decays to that of the disturbing background noise. However, since the minimum is biased towards lower values, an unbiased estimate was obtained by multiplying the local minimum with a bias factor derived from the statistics of the latter [12]. The main drawback of this method is that it takes slightly more than the duration of the minimum search window to update the noise spectrum, when results in delays when tracking a sudden change in the noise power level [13].

In [10], Cohen proposed a new method called minima controlled recursive averaging (MCRA) in which the estimate of the noise is updated by tracking noise-only regions of the noisy speech spectrum over time, which in turn is achieved based on the speech presence probability in each frequency bin. The latter is calculated using the ratio of the noisy speech PSD level to its local minimum over a fixed time window. Then the noise estimate is obtained by averaging past PSD values, with the use of a smoothing parameter which is derived based on the speech presence probability. The main drawback of this method is again the delay in recognizing an abrupt change in the noise level; this delay is almost twice the length of the data window on which the processing is performed [10].

In [11], Cohen proposed a modified version of MCRA called improved minima controlled recursive averaging (IMCRA) [11], aiming at resolving the problems of MCRA. In this method, a different approach is used to track the noise-only regions of the spectrum based on the estimated speech presence probability. The noise estimation procedure includes two iterations of smoothing and minimum tracking. In the first iteration, a rough decision about speech presence probability is made in each frequency bin based on the results of smoothing and minimum tracking. In the second iteration, smoothing in time and frequency is performed which excludes strong speech components in order to boost the efficiency of minimum tracking in speech activity regions [11]. However, since the noise estimate is controlled by minimum tracking, IMCRA still suffers from delays in detecting an increase in the noise level [13].

1.2.2 Speech enhancement methods

Speech enhancement algorithms can be categorized into single-channel and multi-channel algorithms depending on the number of microphones being employed. Single microphone (SM) techniques, which are simple to implement and have lower costs, have been the focus of earlier studies [14] on speech enhancement. In recent years, there have been much interest towards the development of microphone array (MA) techniques, which can coherently process the output of multiple microphones and thereby discriminate sound sources spatially through the applications of beamforming techniques [15]. However those methods are generally have high implementation costs and therefore, there is still a strong interest from industries and academia for improved SM techniques. In this thesis the focus is on SM techniques, and accordingly only these methods are considered in the following literature review.

In general, SM speech enhancement methods can be classified into two main groups. In the first group, the enhancement is done by passing the noisy speech trough an enhancing filter directly in the discrete-time domain. Thus the most critical and challenging issue is to find a proper optimal filter that can remove the noise effectively without making distortions to the speech signal. The optimal filter applied in the time domain should be designed on a short-time basis due to the fact that the speech is highly nonstationary. The procedure is to first divide the speech signal into short-time frames, where the frame length is a few tens of milliseconds. Afterwards, for each of the frames where the speech is now considered to be stationary, the optimal filter is constructed. By passing the noisy speech frame through the constructed filter, the estimate of the clean speech is obtained. However, this method is computationally expensive as it often involves the computation of a matrix inverse [4]. Examples of such processing includes linear convolution and Kalman filtering [16], [17], [18].

In the second group, after decomposing the noisy speech into successive analysis frames, a transform is applied to the windowed frame to produce transform coefficients, and then the enhancement is performed by modifying each coefficient separately. The transform has several advantages as it can act as a decorroletor where the transform coefficients are uncorrelated or even statistically independent. Therefore, the processing operation such as excluding a noisy transform coefficient, can be done on each coefficient separately [19]. One of the most popular transforms is the short-time Fourier transform (STFT) [1], which is used to map the speech samples from a given frame into the frequency domain. The enhancement is performed by modifying STFT coefficients which are converted back to the time-domain using an inverse STFT. These methods, known collectively as frequency domain methods in the literature, are further discussed below. Many other types of transforms have also applied for the purpose of enhancing speech signals in a transform domain. Examples include the subspace methods which apply Karhunen-Loeve Transform (KLT) on each frame of the noisy speech [20], [21], [22] as well as methods which are based on the discrete cosine transform (DCT) and the wavelet transform domains [23],[24], [25] [26].

Generally, it is more practical to process the speech signal in the frequency domain since the vocal tract produces signals based on filtering mechanisms that which can be analyzed or processed more easily in the spectral domain rather than the time domain [1]. In order to process the signals in the STFT domain, the fast Fourier transform (FFT) is usually employed in system implementations. The complete procedure can be explained in four steps as follows [4]:

- As in time domain processing, the noisy speech is divided into short-time frames that overlap partly.
- A tapering window is applied to the speech samples in each frame, which are then mapped to the frequency domain via the FFT.
- To obtain and estimate of the clean speech, an enhancing filter (taking the form of frequency dependent gains) is applied to the complex STFT coefficients.
- Finally, An inverse FFT is applied to the modified STFT coefficients and the enhanced speech is obtained via an overlap-add operation in the time-domain.

This frequency-domain approach is more efficient than its time domain counterpart, due to the use of the computationally efficient FFT algorithm. In addition, because of the decorrelating nature of the STFT, the different complex STFT coefficients can be processed independently, i.e. without any coupling between them. This gives us more flexibility in implementation and in general, results in improved speech enhancement performance [4].

Examples of such STFT-based frequency domain methods include spectral subtraction [27], [28], Wiener filtering [29] and Bayesian approaches [30],[31],[32]. In the spectral subtraction approach, the attempt is to estimate the spectral amplitude (i.e. magnitude of the corresponding STFT coefficient) of the clean speech, from the observed noisy speech. This is mainly done by subtracting an estimate of the noise spectral amplitude from that of the observed noisy speech. Finally, the estimated amplitude is combined with the phase of the noisy speech to produce the desired estimate of the clean speech STFT. In the Wiener filtering approach, the estimate of the clean speech STFT is obtained using a MMSE estimator, where the statistical distributions of the speech and noise are considered to be Gaussian. Similar to the spectral subtraction method, the phase of the clean speech estimate is obtained from that of the noisy speech. Both spectral subtraction and Wiener filtering methods, suffer from the a musical noise which results from the process of obtaining the enhanced speech.

In this thesis, we focus on a group of algorithms, called Bayesian estimators, which fall in the category of frequency domain, single-channel speech enhancement methods. In these estimators, the estimate of the clean speech is obtained by minimizing the expected value of a cost function which provides a measure the error between the estimated and the real speech. It is shown in [33] that the performance of Bayesian estimators is subjectively superior than many other speech enhancement methods. These methods further reviewed below.

Bayesian estimators typically operate in the frequency domain, where the estimate of the clean speech is obtained by modifying the complex STFT coefficients of the speech signal in a given analysis frame of noisy speech.

formulated as estimating the complex STFT coefficients of the speech signal in a given analysis frame of noisy speech. However, it has been shown in [34] and [35] that the spectral amplitude of the speech signal is more relevant than its phase. Therefore, it is more useful to estimate the STSA of the speech signal instead of its STFT coefficients. In such systems the STSA of the speech signal is therefore estimated and then combined with the short-term phase of the observed noisy speech in order to build the enhanced signal.

As explained above, in the Bayesian estimators scheme, the estimate of the clean speech is obtained by minimizing the expected value of a cost function which represents the error between the estimated and the real speech. The performance of these enhancement methods mainly depends on the choice of this cost function as well as certain statistical properties of the speech and noise signals. It is shown in [30] that it is practical to model the STFT coefficients as independent zero-mean complex Gaussian random variables with time-varying variances. All of the algorithms described below use this type of model for the speech and noise signal statistics.

In [30], Ephraim and Malah introduced a well-known Bayesian estimator, known as an MMSE STSA estimator in which the cost function is the mean squared error between the estimated and the true speech STSA under the Gaussian assumption [30]. This approach led to great improvement in speech enhancement performance, specially due to its lower residual noise when compared to the Wiener filter [2]. Subsequently other Bayesian estimators were developed by generalizing MMSE STSA method.

Based on the idea that the human auditory system performs a logarithmic compression of the STSA, Ephraim and Malah proposed an improved version of the MMSE STSA method in [31] which is called log-MMSE. In this method the distortion measure is based on the mean-square error of the log-spectra. The superiority of this method compared to the original MMSE STSA, is in producing lower level of residual noise without introducing additional distortion to the speech signal [31].

Instead of log-MMSE, other estimators have been developed by choosing cost functions that takes into account the internal mechanisms of the human auditory systems. Examples are given by [36] and [37], where masking thresholds are introduced in the the cost function, and in [32] where the cost function is based on perceptual distortion measures.

One of the best cost functions is the weighted Euclidean (WE) measure, introduced in [32], in which the error between the enhanced and clean speech STSA is weighted by the STSA of clean speech raised to a power p. This choice was motivated based on the masking property of the human auditory system, where noise near spectral peaks is more likely to be masked and therefore less audible [32]. The resulting speech enhancement algorithm is referred to as WE in the literature.

Another modified version of the MMSE STSA called β-SA is proposed in [38]. In the underlying cost function, a power law with exponent β, is applied to the square root of the estimated and clean speech. The exponent β is used to avoid over reduction of the noise and better control of the speech distortion.

The Bayesian estimator utilized in this thesis is the modified version of MMSE STSA method, called the Wβ-SA method, recently proposed by Plourde and Champagne in [39]. The cost function used in Wβ-SA generalizes the one used in the two previously proposed methods [32] and [38]. The parameters which are used to build the cost function in Wβ-SA, basically combine those in [32] and [38]. However, these parameters are chosen based on the characteristics of the human auditory system, such as the compressive nonlinearities of the cochlea, the perceived loudness and the ear's masking properties. Choosing the model parameters in this way, decreases the processing gain at high frequencies which in turn provides more noise reduction as well as limiting the speech distortion at lower frequencies. A more detailed technical description of the family of MMSE STSA Bayesian algorithms will be given in Chapter 2.

1.2.3 Data driven speech enhancement methods

Other more sophisticated methods have also been developed in which data-driven statistical learning is applied to derive *a priori* knowledge of the speech and noise descriptors. This knowledge can be used to develop a probabilistic model of the observed data which, in turn, can be employed to derive estimators of the relevant speech and noise statistics. For instance, the obtained *a priori* knowledge can be used to define specific probability density functions (PDF) for the speech and noise spectral components. As an example, the speech PDF can be described using a Laplacian density while the noise PDF can be assumed to be Gaussian [40]. From there, various estimation principles, such as maximum likelihood (ML) or minimum mean square error (MMSE), can be applied to derive the estimates of the unknown noise parameters. Typical methods within this category include the ones based on hidden Markow model (HMM) and linear predictive codebook, which are further described below.

In [41], the parameters of the speech and noise spectral shapes, specifically the auto-regressive (AR) coefficients and associated excitation variances, are modeled using HMMs. This type of modeling is based on multiple hidden states with observable outputs, the states being connected with the transition probabilities of a Markov chain. The HMMs parameters are estimated beforehand, i.e. trained based on data derived from various selected noise types; once the model has been trained, it can applied to noisy speech to derive estimates of the speech and noise AR parameters. In [41], to optimize system performance, the estimated noise variance is scaled by a so-called gain adaptation mechanism, which adjusts the noise level based on processing the data observed during silence regions (non-speech). The AR parameters of the noise model based on the trained HMM are combined with those of the clean speech to obtain an MMSE estimate of the clean speech, as a weighted sum of MMSE estimators corresponding to each state of the HMM for the clean speech signal. In the presence of a stationary background noise, this HMM based method can estimate the noise spectral shape effectively. However, its main problem is that it can only update the noise parameters during non-speech activity periods, and it is therefore slow in adapting to changes in the noise background. Actually, as pointed out in [40], the adaptation speed is comparable to that of the long-term estimate based on minimum tracking in [9]. Another limitation of this HMM based method is that its performance will be degraded when the characteristics of the actual noise differ significantly from those of the noise data used to train the HMMs.

Other examples of such model based systems, are the methods which use trained codebooks of speech and noise LP coefficients to provide the *a priori* information needed in the process of noise statistics estimation. In contrast to HMM based methods which include the excitation variances in the *a priori* information, here the gains are assumed to be unknown and need to be evaluated. Examples of such methods are presented in [42], [43] and [44], which are briefly reviewed below.

In [42], for each pair of speech and noise codebook entries, the speech and noise excitation variances that maximize the likelihood function are computed. Afterwards, the computed excitation variances along with the LP coefficients stored in each pair of speech and noise codevectors are applied to model the speech and noise power spectrum. A log-likelihood score between the observed noisy speech and the modeled one is defined and the estimates of speech and noise spectra, that is the pair of speech and noise codebook which maximize the identified likelihood score, together with the related excitation variances are obtained, corresponding to a standard ML estimation. In [43], the same approach is followed, but a different distortion measure is used instead of the log-likelihood. Indeed it is proved in [43] that maximizing the log-likelihood in equivalent to minimizing the Itakura-Saito measure. Based on this idea, a search is performed through the speech and noise codebooks in order to find the excitation variances which minimize the Itakura-Saito measure. In [44] a further processing step is added to the ML estimation, in order to make the parameter estimation more robust. In this approach, the PDF of the observed noisy speech is defined using the ML estimates of speech and noise. Afterwards, this knowledge of observed data PDF is applied in a MMSE approach, in which the MMSE estimates of the speech and noise LP coefficients along with their excitation variances are derived. This method will be used in this thesis to derive the statistics of the noise. it will therefore be explained in further detail in Chapter 3.

1.3 Thesis Contribution

As discussed before, Wβ-SA method of speech enhancement as demonstrated in [39], shows improved performance compared to other Bayesian speech enhancement methods. However, the results presented in [39] have been obtained under stationary noise conditions, where the required statistics of the noise are obtained beforehand by processing a sample of the clean noise signal. But in practice, we can hardly proceed in this way since the clean noise is not readily available. The other problem is that in reality, the noise which degrades the speech signal quality is nonstationary and its statistics (e.g. spectral properties) change over time.

In this thesis, to overcome this limitation, our main goal is to use one of the data driven methods explained in Section 1.2.3 to derive the statistical knowledge of the noise signal. Once an estimate of the noise statistics is obtained, it will be applied in the Wβ-SA speech enhancement method described in Section 1.2.2 in order to obtain the estimate of the clean speech signal, even in the presence of the noise with nonstationary properties.

The model based method used in this thesis is a combination of the methods proposed in [42] and [44]. Each of these methods exploit trained codebooks of speech and noise LP coefficients to model the required *a priori* knowledge. First, the maximum likelihood estimates of the speech and noise excitation variances are derived using the method proposed in [42]. Then the ML estimates are used in the MMSE approach explained in [44] in order to obtain the final speech and noise LP coefficients and excitation variances. Afterwards, the speech and noise spectra are modeled using the derived parameters. The estimated speech and noise PSDs are then fed into the Wβ-SA speech enhancement scheme to derive the estimate of the clean speech. Since the estimate of the noise is constantly updated, this method performs efficiently in nonstationary environments.

The speech enhancement method used in this work, is the Wβ-SA method developed in [39]. As it was discussed in Section 1.2.2, this method offers a better trade off between noise reduction and speech distortion results by making use of perceptually adjusted parameters. In this thesis, we examine in detail the incorporation of the above codebook based noise estimation method [44] within the Wβ-SA speech enhancement method [39].

This combination is achieved by replacing the noise variance in the calculation of the *a priori* and *a posteriori* SNR parameters, which are then used in the calculation of the gain function. The latter is then applied to the STSA of the observed noisy speech, in order to derive the clean speech data, as will be further explained in Chapter 3.

In Chapter 4, we evaluate the performance of the resulting speech enhancement algorithm which combines the codebook-based scheme with Wβ-SA speech enhancement method. In particular, its performance is compared to that of the STFT-based Wiener filtering method [29] under non-stationary noise conditions. To this end, different types of noise are used, including train, street, car, restaurant and airport noise. The comparison is made by computing PESQ objective measures of speech quality. The results, which are also supported by informal listening, point to the superiority of the newly developed approach over the Wiener filter in terms of both subjective and objective measures.

1.4 Organization

In Chapter 2, various important noise estimation algorithms are first reviewed where we point out the advantages and drawbacks of each technique. Afterwards, the MMSE STSA Bayesian speech enhancement method is explained in detail, followed a presentation of its by the improved versions including Wβ-SA. In Chapter 3, the codebook based parameter estimation method [44] is presented in detail and then it is explained how it can be incorporated within the Wβ-SA speech enhancement method. The performance of the method with respect to different parameter settings and under different noise environment is studied Chapter 4, where objective, i.e. numerical evaluation results are presented. Concluding remarks and possible opportunities for future work are summarized in Chapter 5.

Chapter 2

Background Material

This chapter includes two main sections. In the first section, selected methods of noise PSD estimation which fall into the category of soft-decision approaches are described in detail. In the second section, several speech enhancement algorithms within the category of frequency domain Bayesian STSA approaches are explained, including the Wβ-SA method which plays a central role in this thesis. In our presentation, we try to explain the advantages and drawbacks of the various methods and algorithms under consideration.

2.1 Noise PSD Estimation

As explained before in Section 1.2.1, the soft-decision noise PSD estimation methods differ from the hard-decision ones in the underlying approach used for updating the noise statistics estimates. While these estimates are updated only during silence regions in the hard-decision methods, they are updated continually, i.e. regardless of whether speech is present or absent, in the soft-decision schemes. In this section two noise PSD estimation methods which fall into the category of soft-decision methods are reviewed and their operation is explained. The first method is that of minimum tracking proposed by Martin [9], while the second method is the so-called IMCRA proposed by Cohen [11]. Before proceeding however, we introduce certain modeling elements which are common to both methods.

The general model used in these selected methods in order to represent the discretized noisy speech, is the basic additive noise model, which can be expanded as follows:

$$y(n) = x(n) + w(n) \tag{2.1}$$

where $y(n)$, $x(n)$ and $w(n)$ denote the samples of the noisy speech, the desired speech and the additive noise data respectively, and integer n represents the discrete-time index, where uniform sampling at a given rate F_s is assumed.

In a short observation interval of about 20-40ms, it can be assumed that the desired speech signal $x(n)$ and additive noise $w(n)$ are realizations of independent, zero mean and wide-sense stationary random processes. Therefore, it is useful to separate the set of observed noisy speech samples $y(n), \quad 0 \leq n \leq L$, into overlapping frames with duration less than 40 ms [2]. This can be written as follows:

$$y_l(n) = y(n + \ell M), \qquad 0 \leq n < N, \qquad 0 \leq l < N_f \tag{2.2}$$

where ℓ denotes the frame index, M is the frame advance, N is the frame length with $N \geq M$ ($N - M$ is the number of samples that overlap between two successive frames) and N_f is the total number of frames. An analysis window $h_a(n)$ is applied on each frame for the purpose of trading-off between resolution and the sidelobe suppression in the frequency analysis [2]. Afterwards, each windowed frame of noisy speech data is transformed into the frequency domain using the discrete Fourier transform (DFT) as follows:

$$Y(k, \ell) = \sum_{n=0}^{N-1} y_l(n) h_a(n) e^{-j\frac{2\pi}{N}kn} \tag{2.3}$$

where $k \in \{0, 1, ..., N-1\}$ is the frequency index and $Y(k, \ell)$ denotes the corresponding STFT coefficient of the noisy speech for the lth frame. Therefore, the additive noise model (2.1) can be represented in the STFT domain as:

$$Y(k, \ell) = X(k, \ell) + W(k, \ell) \tag{2.4}$$

where $X(k, \ell)$ and $W(k, \ell)$ denote the STFT coefficients of the clean speech and noise in the lth frame, respectively.

In the literature an speech enhancement, noise estimation refers to the estimation of the variance of $W(k,\ell)$ which under the zero-mean assumption is given by

$$\sigma_W^2(k,\ell) = E\{|W(k,\ell)|^2\}. \tag{2.5}$$

This quantity is also referred to as the short-term power spectrum. Similarly, we can define:

$$\sigma_X^2(k,\ell) = E\{|X(k,\ell)|^2\} \tag{2.6}$$
$$\sigma_Y^2(k,\ell) = E\{|Y(k,\ell)|^2\}. \tag{2.7}$$

Under the independence assumption it follows from (2.4) that:

$$\sigma_Y^2(k,\ell) = \sigma_X^2(k,\ell) + \sigma_W^2(l,\ell) \tag{2.8}$$

The main goal of the methods reviewed in the following sub-sections is to obtain a running estimate of the noise PSD, i.e. $\sigma_W^2(k,l)$ in (2.5), based on the observations of the noise speech STFT $Y(k,\ell)$.

2.1.1 Minimum statistics (MS) noise estimation

In [9], Martin proposed an original method for estimating the noise PSD from the observed noisy speech. This method, which is based on minimum statistics and optimal smoothing, relies on two fundamental premises. First, it is assumed that the clean speech and additive noise signals are statistically independent. Second, as it is observed experimentally, the PSD level of the noisy speech signal often decays to that of the background noise. Therefore, the estimate of the noise PSD can be derived by tracking the minimum of the noisy speech power spectrum.

An estimate of the noise PSD $\sigma_W^2(k,l)$ in (2.5) can be obtained through a first order recursive averaging of the instantaneous magnitude spectrum $|Y(k,\ell)|^2$, also called periodogram , as follows:

$$P(k,\ell) = \alpha P(k,\ell-1) + (1-\alpha)|Y(k,\ell)|^2 \tag{2.9}$$

where $P(k,\ell)$ is the desired estimate and $0 \le \alpha \le 1$ is a smoothing parameter.

More generally, the smoothing parameter α used in (2.9) can be considered as time and frequency dependent. i.e. $\alpha \equiv \alpha(k, \ell)$. Using such a time and frequency dependent parameter, (2.9) can be rewritten as follows:

$$P(k, \ell) = \alpha(k, \ell) P(k, \ell - 1) + (1 - \alpha(k, \ell))|Y(k, \ell)|^2 \tag{2.10}$$

In order to derive an optimal value for $\alpha(k, \ell)$, only the speech silence regions are considered. Since in theory the speech signal PSD is equal to zero during these intervals, i.e. $\sigma_S^2(k, \ell) = 0$, $P(k, \ell)$ should be as close as possible to the noise PSD. This can be fulfilled by minimizing the mean squared error between $P(k, \ell)$ and $\sigma_D^2(k, \ell)$, given the previous estimate $P(k, \ell - 1)$, which can be formally expanded as:

$$E\{(P(k, \ell) - \sigma_W^2(k, \ell))^2 | P(k, \ell - 1)\}. \tag{2.11}$$

Substituting (2.9) in (2.11) and setting the first derivative with respect to α to zero, the optimal value of α, denoted as α_{opt} is derived as follows:

$$\alpha_{opt}(k, \ell) = \frac{1}{1 + (P(k, \ell - 1)/\sigma_W^2(k, \ell) - 1)^2}. \tag{2.12}$$

In practical implementations, it has been observed that the use of (2.12) leads to errors in estimating the noise PSD, and therefore the smoothing parameter should be modified. To this end, the estimation errors are tracked by comparing $P(k, \ell)$ and a reference quantity, which is considered to be the frequency averaged periodogram. Specifically an error monitoring algorithm is employed in [9] which compares the average smoothed PSD estimate of the previous frame $(1/N) \sum_{k=0}^{N-1} P(k, \ell - 1)$ and the average periodogram of the current frame $(1/N) \sum_{k=0}^{N-1} |Y(k, \ell)|^2$, where the average is over all the frequency bins. A correction factor denoted as $\alpha_c(l)$ is calculated in each frame using the ratio of these averaged quantities. The corrected value of the optimal smoothing parameter is calculated by multiplying the right hand side of (2.12) by $\alpha_c(l)$ as in:

$$\alpha_{opt}(k, \ell) = \frac{\alpha_{max} \alpha_c(\ell)}{1 + (P(k, \ell - 1)/\sigma_W^2(k, \ell) - 1)^2} \tag{2.13}$$

where $\alpha_{max} = 0.96$ is used as an upper limit on the smoothing parameter.

In the minimum tracking PSD estimation approach, the minimum value of the smoothed PSD estimate $P(k, \ell)$ (2.10) over a finite temporal window of length L frames is used as the desired estimate, that is:

$$P_{min}(k, \ell) = min\{P(k, m) : L - l < m \leq L\} \tag{2.14}$$

The minimum noise PSD estimate is necessarily biased since, for nontrivial probability densities, the minimum value of a set of random variables is smaller than their mean [45]. Therefore, a bias factor is needed to compensate the use of the minimum operation in this estimation approach. Similar to the procedure for finding the smoothing parameter, one only needs to consider the speech silence periods of the noisy speech in order to derive the bias factor.

It can be seen from the denominator of the optimum smoothing parameter in (2.13) that the PSD estimate $P(k, \ell)$ is normalized by the variance of the noise or PSD, i.e. $\sigma_W^2(k, \ell)$. More generally, it can be shown that the PDF of $P(k, \ell)$ in (2.10) in scaled by $\sigma_W^2(k, \ell)$, which in turn implies that the minimum statistics of the smoothed PSD $P(k, \ell)$ is also scaled by $\sigma_W^2(k, \ell)$. Therefore, it can be concluded that the mean and the variance of $P_{min}(k, \ell)$ are respectively proportional to $\sigma_W^2(k, \ell)$ and $\sigma_W^4(k, \ell)$, and without loss of generality, it is sufficient to compute the mean and the variance for the case that $\sigma_W^2(k, \ell) = 1$ [9]. Hence, the bias compensation factor can be defined as:

$$B_{min}^{-1}(k, \ell) = E\{P_{min}(k, \ell)\}_{\sigma_W^2(k,\ell)=1} \tag{2.15}$$

and after some manipulations, it is obtained as follows:

$$B_{min}(k, \ell) \approx 1 + (L - 1)\frac{2}{Q_{eq}}(k, \ell) \tag{2.16}$$

where L is the window length and Q_{eq} is the so-called equivalent degrees of freedom, defined as $Q_{eq}(k, \ell) = 2\sigma_W^4(k, \ell)/\mathrm{var}\{P(k, \ell)\}$.

Finally, the desired unbiased noise PSD estimate can be expressed as follows:

$$\hat{\sigma}_W^2(k, \ell) = B_{min}(k, \ell)P_{min}(k, \ell) \tag{2.17}$$

It can be seen that the minimum tracking of the noise PSD is done over a fixed window of length L. The window length must be large enough to cover at least one silence period as well as extending beyond the broadest peaks of speech energy in the noisy speech waveform [12]. As we are dealing with sources of nonstationary noise, there may also be abrupt changes in the noise level. An example is a talker who uses his cell phone while moving from a quiet place to a noisy one [12]. Since the minimum is obtained over a fixed window of length L, it may take a time in excess of L frames to track a sudden change in noise. This is clearly an undesirable behavior, since we are looking for a system which can track the nonstationarities of the noise source in real time. Other advanced methods have been proposed to reduce this delay in tracking an abrupt change in the noise level in [10] and [11].

2.1.2 Minima controlled recursive averaging (MCRA)

In [10], Cohen presented a more sophisticated noise PSD estimation method, in which the noise estimate is updated by averaging the past spectral values of the noisy speech. The estimate is controlled by a smoothing parameter, which is dependent on both time and frequency. This dependence is achieved via the *a priori* speech absence probability in each frequency bin separately. In order to derive the speech presence probability, first the local minimum is obtained over a fixed window of time, and then the probability is calculated using the ratio of the noisy speech power spectrum to the local minimum in that frame. This method, called Minima Controlled Recursive Averaging (MCRA) is explained in further detail below:

The noise estimation process is mainly based on two hypotheses, corresponding on whether the speech is present or not. Specifically, these hypotheses can be expressed as follows:

$$\begin{aligned} H_0(k,\ell) &: \quad Y(k,\ell) = W(k,\ell) \\ H_1(k,\ell) &: \quad Y(k,\ell) = X(k,\ell) + W(k,\ell) \end{aligned} \tag{2.18}$$

where $Y(k,\ell)$, $S(k,\ell)$ and $D(k,\ell)$ denote the STFT coefficients of the noisy speech, clean speech and noise, l denotes the frame index and k represents the frequency bin. As can be observed from the definitions of the hypotheses, it is considered under $H_0(k,\ell)$ that the speech is absent, while under $H_1(k,\ell)$ the presence of both speech and noise is assumed.

Let $P(k,\ell)$ denote the desired estimate of the noise spectrum (or variance), i.e. $\sigma_W^2(k,\ell) = E\{|W(k,\ell)|^2\}$. Based on the above hypotheses, $P(k,\ell)$ is updated differently during frames where speech is absent and frames where it is present, as follows:

$$\begin{aligned} H_0(k,l) &: P(k,\ell+1) = \alpha_d P(k,\ell) + (1-\alpha_d)|Y(k,\ell)|^2 \\ H_1(k,l) &: P(k,\ell+1) = P(k,\ell) \end{aligned} \tag{2.19}$$

where $\alpha_d \quad (0 < \alpha_d < 1)$ is the smoothing factor. It can be observed that the noise spectrum estimate is only updated during frames where speech is absent. The two equations in (2.19) can be combined in a single equation, by making use of the speech presence probability as follows:

$$P(k,\ell+1) = P(k,\ell)p_r(k,\ell) + [\alpha_d P(k,\ell) + (1-\alpha_d)|Y(k,\ell)|^2](1-p(k,\ell)) \tag{2.20}$$

where $p_r(k,\ell)$ denotes the conditional speech presence probability under the observation of $Y(k,\ell)$, defined as $p_r(H_1(k,\ell)|Y(k,\ell)$. Equivalently, recursion (2.20) can be written in the form:

$$P(k,\ell+1) = \tilde{\alpha}_d(k,\ell)P(k,\ell) + [1-\tilde{\alpha}_d(k,\ell)]|Y(k,\ell)|^2 \tag{2.21}$$

where the new smoothing parameter $\tilde{\alpha}_d(k,\ell)$ is given by:

$$\tilde{\alpha}_d(k,\ell) = \alpha_d + (1-\alpha_d)p_r(k,\ell) \tag{2.22}$$

As can be observed from (2.22), the computation of $\tilde{\alpha}_d(k,\ell)$ requires the knowledge of the conditional speech presence probability $p_r(k,\ell)$. In [10], the latter is estimated based on the ratio between the local energy of the noisy speech and its minimum over a specified time window of length L frames, where the value of L is chosen to cover the duration of the broadest peaks in speech activity. The local energy of the noisy speech is obtained by smoothing the squared magnitude spectrum of the noisy speech in the time and frequency domain, as explained below.

In the frequency domain, the smoothed power spectrum is obtained as follows:

$$S_f(k,\ell) = \sum_{-w}^{w} b(i)|Y(k-i,\ell)|^2 \tag{2.23}$$

where $b(i)$ is a window function whose length is $2w+1$. In the time domain, the smoothing is performed by a first order recursive averaging as follows:

$$S(k,\ell) = \alpha_s S(\ell-1,k) + (1-\alpha_s)S_f(k,\ell) \tag{2.24}$$

where α_s is a smoothing factor. The minimum value of a specific frame is obtained by comparing the local energy in that frame to the minimum value of the previous frame, as in:

$$S_{min}(k,\ell) = \min\{S_{min}(k,\ell-1), S(k,\ell)\} \tag{2.25}$$

The actual implementation of this scheme is slightly more complicated than this as it involves the use of temporary variable, say $S_{tmp}(k,\ell)$, which is employed and re-initialized with every block of L consecutive frames. We refer to reader to [10] for additional detail.

Let the indicator function $I(k,\ell) \in \{0,1\}$ represent the presence of speech in each frame, that is $I(k,\ell) = 1$ when speech is present and 0 otherwise. $I(k,\ell)$ is specified by using the ratio between the noisy speech power and its minimum, defined as $S_r(k,\ell) = S(k,\ell)/S_{min}(k,\ell)$. Let $\delta > 0$ be a threshold introduced to determine whether the speech is present or absent in a given frame. Then we have:

$$I(k,\ell) = \begin{cases} 1 & \text{if } S_r(k,\ell) > \delta \quad \text{(speech present)} \\ 0 & \text{if } S_r(k,\ell) < \delta \quad \text{(speech absent)} \end{cases} \tag{2.26}$$

Using this indicator function, the speech presence probability is estimated as:

$$p_r(k,\ell) = \alpha_p p_r(k,\ell-1) + (1-\alpha_p)I(k,\ell) \tag{2.27}$$

where $0 < \alpha_p < 1$ is a smoothing parameter.

Finally having estimated the conditional speech presence probability $p_r(k,\ell)$ as above, the smoothing parameter $\tilde{\alpha}_d$ is computed and the noise PSD estimate $P(k,\ell)$ is updated using (2.21). As pointed out above, however, because of a time window of L frames, this proposed approach also has a memory in excess of L frames. Consequently, a similar problem as in [9] appears in tracking sudden change in the noise power.

Improved version of MCRA (IMCRA)

In [11], Cohen proposed an improved version of MCRA which is called IMCRA in the literature. Likewise MCRA, it includes averaging past spectral values using a smoothing factor which is dependent on the speech presence probability. Compared to MCRA, improvements have been made with respect to minimum tracking, speech presence probability calculation and introducing a bias compensation factor [12].

Similar to (2.18) in MCRA, two hypotheses $H_0(k,l)$ and $H_1(k,l)$ are defined, which refer to the case of speech absence and speech presence. The noise PSD estimate is updated exactly as in (2.21) and (2.22), where the conditional speech presence probability $p_r(k,\ell) = p_r(H_1(k,\ell)|Y(k,\ell)|)$ is now computed on the basis of a Gaussian model for the speech signal and noise component. Specifically,

$$p_r(k,\ell) = \left\{1 + \frac{q(k,\ell)}{1-q(k,\ell)}(1+\xi(k,\ell))\exp(-\upsilon(k,\ell))\right\}^{-1}, \tag{2.28}$$

where

$$\gamma(k,\ell) = \frac{\sigma_X^2(k,\ell)}{\sigma_W^2(k,\ell)} \qquad \zeta(k,\ell) = \frac{|Y(k,\ell)|^2}{\sigma_W^2(k,\ell)} \tag{2.29}$$

are the *a priori* and *a posteriori* SNRs, $\upsilon = \gamma\zeta/(1+\zeta)$ and $q(k,\ell) = pr(H_0|(k,\ell))$ is the *a priori* probability for speech absence.

The estimation of $q(k,\ell)$ is controlled by the minima values of a smoothed power spectrum of the noisy speech and comprises two iterations of smoothing and minimum tracking, as explained below:

The first iteration provides a coarse decision to identify the frames of speech in each frequency band. At first, the noisy speech PSD $S(k,\ell)$ is smoothed in frequency and time domain as in (2.23) and (2.24). Afterwards, the minimum of the smoothed noisy PSD is searched over a window of length L frames, that is:

$$S_{min}(k,\ell) = min\{S(k,m)|\ell - L < m \leq \ell\} \tag{2.30}$$

Consequently, *a posterior* and *a prior* SNR measures are defined respectively as follows:

$$\gamma_{min}(k,\ell) = \frac{|Y(k,\ell)|^2}{B_{min}S_{min}(k,\ell)} \qquad \zeta(k,\ell) = \frac{S(k,\ell)}{B_{min}S_{min}(k,\ell)} \tag{2.31}$$

where B_{min} denoted the bias of the minimum noise power estimate. Based on these two SNRs, a criterion is defined upon which a decision is made whether the speech is present or not, as follows:

$$I(k,\ell) = \begin{cases} 1 & \text{if } \gamma_{min}(k,\ell) < \gamma_0 \quad \text{and} \quad \zeta(k,\ell) < \zeta_0 \quad \text{speech absent} \\ 0 & \text{otherwise} \quad \text{speech present.} \end{cases} \tag{2.32}$$

where the parameters $\gamma_0 = 4.6$ and $\zeta_0 = 1.67$.

The second iteration of smoothing includes only the power spectral components, which have been identified as containing primarily noise. The smoothing in frequency domain is obtained based on the made decision as follows:

$$\tilde{S}_f(k,\ell) = \begin{cases} \dfrac{\sum_{-w}^{w} b(i) I(k-i,\ell)|Y(k-i,\ell|^2}{\sum_{-w}^{w} b(i) I(k-i,\ell)} & \text{if } \sum_{-w}^{w} I(k-i,\ell) \neq 0 \\ \tilde{S}_f(k,\ell-1) & \text{otherwise} \end{cases} \tag{2.33}$$

Consequently, smoothing in time is achieved by a recursive averaging as follows:

$$\tilde{S}(k,\ell) = \alpha_s \tilde{S}(k,\ell-1) + (1-\alpha_s)\tilde{S}_f(k,\ell) \tag{2.34}$$

The minimum power spectrum $\tilde{S}_{min}(k,\ell)$ of $\tilde{S}(k,\ell)$ is derived in the second iteration of minimum tracking as in (2.30). Similar to the first iteration, SNR measures are computed as follows:

$$\tilde{\gamma}_{min}(k,\ell) = \frac{|Y(k,\ell)|^2}{B_{min}\tilde{S}_{min}(k,\ell)} \qquad \tilde{\zeta}(k,\ell) = \frac{S(k,\ell)}{B_{min}\tilde{S}_{min}(k,\ell)}. \tag{2.35}$$

Finally, the speech absence probability is derived as follows:

$$\hat{q}(k,\ell) = \begin{cases} 1, & \text{if } \tilde{\gamma}_{min}(k,\ell) < 1 \text{ and } \tilde{\zeta}(k,\ell) < \zeta_0 \\ \dfrac{\gamma_1 - \tilde{\gamma}_{min}(k,\ell)}{\gamma_1 - 1}, & \text{if } 1 < \tilde{\gamma}_{min}(k,\ell) < \gamma_1 \text{ and } \tilde{\zeta}(k,\ell) < \zeta_0 \\ 0, & \text{otherwise} \end{cases} \tag{2.36}$$

where $\gamma_1 = 3$ and $\zeta_0 = 1.67$.

Having found the speech absence probability $q(k,\ell)$, it is possible to compute the conditional speech presence probability $p_r(k,\ell)$ (2.28) and then update the noise PSD estimate $P(k,\ell)$ by means of (2.21) and (2.22). The delay of this method is slightly less than MCRA, but remains significant, specially in the case of an abrupt change in the noise power.

2.2 Bayesian Speech Enhancement Algorithms

As discussed in Chapter 1, there is a strong motivation to use single-channel speech enhancement systems in many applications, due to their low cost and small size. Among all the competing methods, frequency domain Bayesian approaches have received considerable attention. In this section, we review several Bayesian approaches for single-channel speech enhancement, including Wβ-SA method which we utilize in this thesis.

In general, we seek to find an estimate of the clean speech spectrum, denoted by $\hat{X}(k,\ell)$, based on the observation of the noisy speech STFT coefficients $Y(k,\ell)$ introduced in (2.4). At first, the estimate of the clean speech spectrum is derived in each frame of the noisy data separately. Afterwards, the time domain estimate of the clean speech in each frame is evaluated using the inverse Fourier transform which we denote as $\hat{X}(k,\ell)$. Then the clean speech estimates from all the frames are combined using an overlap-add method in order to derive the overall time domain estimate of the speech signal [46]. In the sequel, considering that frequency domain processing is done on each frame separately, we shall drop the frame index l and use the notation $\hat{X}_k = X(k,\ell)$, $\hat{Y}_k = Y(k,\ell)$, etc, To simplify the presentation.

In order to estimate $\hat{X}_k$ from Y_k, a distance metric, or cost function is defined between X_k and $\hat{X}_k$. Specifically, in the Bayesian approach, the speech estimate $\hat{X}_k$ is derived by minimizing the expected value of a cost function as a measure of the error between X_k and $\hat{X}_k$, as given by:

$$E\{C(X_k,\hat{X}_k)\} = \int\int C(X_k,\hat{X}_k) f_{X_k,Y_k}(X_k,Y_k) dX_k dY_k \tag{2.37}$$

where $f_{X_k,Y_k}(X_k,Y_k)$ is the joint PDF of X_k and Y_k. Equation (2.37) can be rewritten as in the form:

$$E(C(X_k,\hat{X}_k)) = \int f_{Y_k}(Y_k) \int C(X_k,\hat{X}_k) f_{X_k|Y_k}(X_k|Y_k) dX_k dY_k \tag{2.38}$$

where $f_{X_k|Y_k}(X_k|Y_k)$ is the conditional PDF of X_k given Y_k and is often referred to the *a posteriori* PDF in the literature, while $f_{X_k}(X_k)$ is called *a priori* PDF. From (2.38), we have:

$$E(C(X_k,\hat{X}_k)) \geq \int f_{Y_k}(Y_k) \min_{\hat{X}_k} \left\{ \int C(X_k,\hat{X}_k) f_{X_k|Y_k}(X_k|Y_k) dX_k \right\} dY_k \tag{2.39}$$

Therefore, in order to minimize the expected value of the cost function, it is sufficient to minimize the inner integral in (2.38). Finally, the Bayesian estimate $\hat{X}_k$ is obtained as:

$$\hat{X}_k = \arg\min_{\hat{X}_k} \int C(X_k, \hat{X}_k) f_{X_k|Y_k}(X_k|Y_k) dX_k \tag{2.40}$$

In general, $\hat{X}_k$ includes an the amplitude and phase estimate of the clean speech spectrum. In many applications of speech processing however, it is more relevant to estimate the spectral amplitude of the speech signal rather than its phase [34], [35]. Therefore, Bayesian estimators have been developed in which the spectral amplitude of the STSA of the speech is estimated and then combined with the phase of the noisy speech [30]. The estimates referred to as Bayesian STSA estimation in the literature, can be expanded in the form:

$$\hat{\mathcal{X}}_k^0 = \arg min_{\hat{\mathcal{X}}_k} \int_0^\infty C(\mathcal{X}_k, \hat{\mathcal{X}}_k) f_{\mathcal{X}_k|Y_k}(\mathcal{X}_k|Y_k) d\mathcal{X}_k \tag{2.41}$$

where $\mathcal{X}_k = |X_k|$ denotes the STSA of the clean speech, and $\hat{\mathcal{X}}_k^0$ is the corresponding Bayesian estimate. Having found the amplitude of the enhanced speech spectrum using (2.41), it is combined with the phase of the noisy speech to obtain the final estimate of the clean speech as follows:

$$\hat{X}_k = \hat{\mathcal{X}}_k^0 e^{j\angle Y_k} \tag{2.42}$$

In general, the problem of estimating the clean speech with such Bayesian estimator mainly depends on defining a proper cost function $C(X_k, \hat{X}_k)$ and statistical model for signal and noise components [2], and different Bayesian STSA estimators have been developed in this way. In the following sub-sections we review some of the most important frequency domain Bayesian STSA estimators in detail.

2.2.1 The MMSE STSA estimator

In [30], Ephraim and Malah proposed a Bayesian STSA estimation method, which is known as MMSE estimator in the literature. In the model they proposed, the STFT coefficients of the speech and noise are considered to be independent complex Gaussian random variables. The corresponding marginal PDFs are given by:

$$f_{X_k}(X_k) = \frac{1}{\pi\sigma_{X,k}^2} e^{-|X_k|^2/\sigma_{X,k}^2} \tag{2.43}$$

$$f_{W_k}(W_k) = \frac{1}{\pi\sigma_{W,k}^2} e^{-|W_k|^2/\sigma_{W,k}^2} \tag{2.44}$$

where $\sigma_{X,k}^2 = E\{|X_k|^2\}$ and $\sigma_{W,k}^2 = E\{|W_k|^2\}$ are the speech and noise variances, respectively.

The cost function they proposed is the squared error between the clean speech STSA and its estimate, which can be expanded as:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k - \hat{\mathcal{X}}_k)^2. \tag{2.45}$$

Using (2.45) in (2.41), the corresponding Bayesian STSA estimator can be obtained as:

$$\hat{\mathcal{X}}_k^0 = E\{\mathcal{X}|Y_k\} = \int |X_k| f_{X_k|Y_k}(X_k|Y_k) dX_k \tag{2.46}$$

Using Bayes rule, (2.46) can be rewritten in the form:

$$\hat{\mathcal{X}}_k^0 = \frac{\int |X_k| f_{Y_k|X_k}(Y_k|X_k) f_{X_k}(X_k) dX_k}{\int f_{Y_k|X_k}(Y_k|X_k) f_{X_k}(X_k) dX_k} \tag{2.47}$$

Under the additive noise model (2.4) for the STFT coefficients Y_k, it follows from the above independence assumption that $f_{Y_k|X_k}(Y_k|X_k) = f_{W_k}(Y_k - X_k)$. Hence (2.47) can be further simplified as:

$$\hat{\mathcal{X}}_k^0 = \frac{\int |X_k| f_{W_k}(Y_k - X_k) f_{X_k}(X_k) dX_k}{\int f_{W_k}(Y_k - X_k) f_{X_k}(X_k) dX_k} \tag{2.48}$$

Applying (2.44) in (2.48), and changing coordinates in the complex plane X_k from rectangular to polar , the MMSE estimator is derived and can be written as follows:

$$\hat{\mathcal{X}}_k^0 = G_k|Y_k| \tag{2.49}$$

where $G_k > 0$ is the gain applied to the spectral magnitude of the noisy speech. This gain is defined as follows:

$$G_k = \frac{\sqrt{\pi v_k}}{2\gamma_k}\exp(\frac{-v_k}{2})[(1+v_k)I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})] \tag{2.50}$$

where $I_0(.)$ and $I_1(.)$ are the modified Bessel functions of zero and first order, respectively, and

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k \qquad \xi_k = \frac{\sigma^2_{X,k}}{\sigma^2_{W,k}} \qquad \gamma_k = \frac{|Y_k|^2}{\sigma^2_{W,k}} \tag{2.51}$$

The parameters ξ_k and γ_k are referred to as the *a priori* and *a posteriori* SNRs, respectively.

2.2.2 Improved forms of MMSE STSA

MMSE log-STSA

In [31], Ephraim and Malah proposed a more advanced form of the MMSE STSA estimator, in which the cost function (2.45) is modified. Specifically, based on the observation that the human auditory system performs a logarithmic compression of the STSA [31], the modified cost function uses the logarithm of the STSA, rather than the STSA itself. Hence the MMSE log-STSA estimator is based on the modified cost function:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = [\ln(\mathcal{X}_k) - \ln(\hat{\mathcal{X}}_k)]^2. \tag{2.52}$$

where ln(.) is the natural logarithm in base 2. Applying (2.52) in (2.41), it can be shown that:

$$\hat{\mathcal{X}}_k^0 = \exp(E[\ln \mathcal{X}|Y_k]) \tag{2.53}$$

Using the same statistical model and following the same procedure as in [30], this estimator can be expanded as in (2.49) where the gain function G_k is now given by:

$$G_k = \frac{v_k}{\gamma_k} \exp(\frac{1}{2} \int_{v_k}^{\infty} \frac{e^{-t}}{t} dt) \tag{2.54}$$

Compared to the MMSE STSA estimator, this method usually results in lower residual errors while slightly increasing the speech distortion [2].

β-order STSA MMSE

Another generalized form of the MMSE STSA estimator was introduced in [38] which is called β-order STSA MMSE estimator. The cost function used in the work is defined as:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta})^2 \tag{2.55}$$

where β is a positive real parameter, introduced as a means to trade-off between speech distortion and noise reduction.

Using (2.55) in (2.41), it can be shown that:

$$\hat{\mathcal{X}}_k^0 = \sqrt{\beta} E\{\mathcal{X}_k^{\beta} | Y_k\} \tag{2.56}$$

Following the same steps as before, this estimator can be expanded as in (2.49) where the gain function is evaluated as follows:

$$G_k = \frac{\sqrt{v_k}}{\gamma_k} \left[\Gamma(\frac{\beta}{2} + 1) M(-\frac{\beta}{2}, 1; -v_k) \right]^{1/\beta}. \tag{2.57}$$

In this expansion, $\Gamma(.)$ stands for the gamma function, which is defined as follows:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \tag{2.58}$$

and $M(a, b; z)$ is the confluent hypergeometric function defined as [47]:

$$M(a, b; z) = 1 + \frac{a}{b} \frac{z}{1!} + \frac{a(a+1)}{b(b+1)} \frac{z^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)} \frac{z^3}{3!} + ... \tag{2.59}$$

In general, as the exponent β decreases towards 0, the gain G_k (2.57) increases which results in more noise suppression while producing more speech distortion. In [38] the value of β is adapted in each frame based on the corresponding SNR in that frame, such that a smaller β is assigned to frames with smaller SNR and vice versa. Therefore, more noise is removed in frames with low SNR, which speech distortion is limited in frames with larger SNR.

Weighted euclidean

In [32], Loizou studied the functionality of several perceptually meaningful distance measures such as the weighted likelihood ratio, the Itakura-Saito distance measure and the COSH distance measure. He proposed a cost function based on a perceptually-weighted error criterion, referred to as the weighted Euclidean (WE) estimator in the literature. The WE cost function can be expressed as follows:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \mathcal{X}_k^p(\mathcal{X}_k - \hat{\mathcal{X}}_k)^2 \tag{2.60}$$

where p is a real parameter which is larger than -2. Proceeding as before, the corresponding gain function of the WE estimator is obtained in the form:

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \frac{\Gamma(\frac{p+1}{2}+1)}{\Gamma(\frac{p}{2}+1)} \frac{M(-\frac{p+1}{2}, 1; -\upsilon_k)}{M(-\frac{p}{2}, 1; -\upsilon_k)} \tag{2.61}$$

where the parameters γ_k and υ_k are already defined in (2.51), respectively. Similar to β in [38], the parameter p controls the trade-off between the noise reduction and the speech distortion; in particular, a smaller value of p will contribute to suppress more noise at the expense of more speech distortion. In practice, the value of $p = -1$ offers a good compromise between the speech distortion and noise reduction.

Wβ-SA STSA

In [39], Plourde and Champagne introduced a new family of Bayesian estimators where the cost function includes both a power law and a weighting factor. Specifically, in this new estimator, the cost function is given by

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \left(\frac{\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^\beta}{\mathcal{X}_k^\alpha}\right)^2 \tag{2.62}$$

where p is the related to the parameter p in the WE estimator [32] through $\alpha = -p/2$ and β is related to the β-order STSA estimator [38]. In the Wβ-SA estimator, these parameters are chosen based on the human auditory systems and ear's masking properties; as a result they become frequency dependent. This characteristic of the proposed Wβ-SA estimator results in a better noise reduction while controlling the speech distortion.

Using the cost function defined in (2.62) and proceeding as the other Bayesian STSA estimators, it can be shown in [39] that the gain for this estimator is:

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \left(\frac{\Gamma\left(\frac{\beta}{2} - \alpha + 1\right) M\left(\alpha - \frac{\beta}{2}, 1; -\upsilon_k\right)}{\Gamma\left(-\alpha + 1\right) M\left(\alpha, 1; -\upsilon_k\right)} \right)^{1/\beta} \tag{2.63}$$

where $\Gamma(a)$ and $M(a, b; z)$ are the gamma and confluent hypergeometric functions, $\beta > 2(\alpha - 1)$, $\alpha < 1$, $\upsilon_k = \gamma_k \xi/(1 + \xi_k)$, with γ_k and ξ_k being *a posteriori* and *a priori* SNR, respectively, as defined in (2.51).

In [39], it is also shown that appropriate values of β and α can be chosen by considering the human auditory system rather the frame SNR. Considering the perceived loudness of sound and the compressive nonlinearities of the cochlea, it is suggested that β be chosen for each frequency bin as follows:

$$\beta_k = d_k \frac{(\beta_{\text{high}} - \beta_{\text{low}})}{\frac{1}{\rho} \log_{10}\left(\frac{F_s}{2A} + 1\right)} + \beta_{\text{low}}, \tag{2.64}$$

where $\beta_{\text{low}} = 1, \beta_{\text{high}} = 0.2$, F_s is the sampling frequency and d_k is defined as follows:

$$d_k = \frac{1}{\rho} \log_{10}\left(\frac{f_k}{A} + 1\right) \tag{2.65}$$

where $\rho = 0.06$ and A is a scaling parameter allowing for the frequency f_k to be expressed in Hz. Considering the masking properties of the human auditory system, it is shown that α can also be calculated in a frequency dependent measure given by

$$\alpha_k = \begin{cases} \alpha_{\text{low}} & f_k \leq 2 \text{ kHz} \\ \dfrac{(f_k - 2000)(\alpha_{\text{high}} - \alpha_{\text{low}})}{F_s/2 - 2000} + \alpha_{\text{low}} & \text{else} \end{cases} \tag{2.66}$$

where $\alpha_{\text{low}} = 0.5$ and $\alpha_{\text{high}} = 0.9$.

Chapter 3

Combination of Speech Enhancement and Noise Estimation Algorithms

In the first part of this chapter, a codebook based approach for estimating noise and speech statistics, specially the short term predictor (STP) parameters, is presented. In the second part, the incorporation of the codebook based method with the Wβ-SA speech enhancement algorithm is explained in detail.

3.1 Codebook Based Noise PSD Estimation

As exposed in Section 1.2.3, an important category of speech enhancement methods employs *a priori* knowledge of speech and noise to estimate the statistics of the noise signal. In particular, the method which is presented in this section, exploits trained codebooks of speech and noise LP coefficients to provide the required *a priori* knowledge.

In general, the possible shapes of the speech spectral envelop are constrained due to physiology of speech production. As will be further explained in Section 3.1.1, one way to specify the spectral envelop of a signal is by using its LP power spectrum [40], where the spectral envelop is dependent on the LP coefficients. The possible spectral shapes for the speech and noise signals can be modeled using sufficiently large codebooks of speech and noise LP coefficients, obtained from large training data sets. Such trained codebooks are then used as the *a priori* information on the speech and noise signals, which can be exploited in various applications of speech processing, to estimate the statistics of the speech and noise signals.

The codebook based method used in this thesis, employs a Bayesian MMSE approach for the estimation of the short-term predictor parameters of speech and noise. The short-term predictor parameters mainly refer to the AR coefficients and the excitation variance (gain). In other words, the MMSE estimates of the speech and noise AR spectra are derived, which mainly include the estimation of AR coefficients and the excitation variances, that define the AR spectra.

Given the observation of the noisy data, there are two possible avenues for the estimation of unknown speech and noise parameters. If the parameters are assumed to be deterministic (but unknown), the estimation process is counted as a classical estimation, including ML estimation. On the other hand, if the unknown parameters are assumed to be random variables with a specific joint PDF, then this estimation is called Bayesian estimation. [40] . In [43], Srinivasan and Kleijn proposed a codebook based method based on the first concept, while in [44], Srinivasan, Samuelsson and Kleijn introduced another method following the second approach. The method used in this thesis, is a combination the of methods presented in [42] and [44]. The two methods are explained in Sections 3.1.3 and 3.1.4, respectively, but first the AR model of the speech and noise PSD is explained in Section 3.1.1, while in Section 3.1.2 we briefly review the Lloyd algorithm for codebook generation.

3.1.1 Autoregressive modeling of speech spectra

In general, the LP coding is widely used in speech processing applications since it provides an accurate and economical representation of relevant speech parameters that can reduce transmission rates in speech coding and lead to efficient speech synthesis [1]. LP parameters provide a rigorous representation of the speech spectral magnitude, while it remains relatively simple in terms of computation. The main applications of LP includes low-bit rate speech coding, adaptive digital filters and speech recognition systems [1].

Based on the human speech production system, the generation of each phoneme of speech involves two factors, i.e. the source excitation and the vocal tract shaping. Considering these two factors, the speech production system can be modeled as is shown in Figure 3.1, where the vocal tract, modeled as a linear filter with the impulse response $h(n)$, is excited by a discrete time glottal signal $u(n)$ to produce the speech signal $x(n)$. In case of unvoiced sounds, where the excitation is similar to white noise, $u(n)$ is chosen to have a flat spectrum. For voice sounds, the source is modeled as periodic impulse train with period N samples, where N is selected for the pitch period [1].

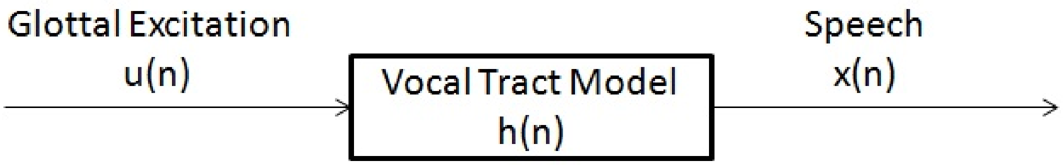


Fig. 3.1 Speech production system

A standard model for the vocal tract filter, is the autoregressive moving average (ARMA) model. In this case, the speech sample $x(n)$ is formed as a linear combination of past outputs and the present and past inputs. This can be expanded as follows:

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + G \sum_{l=0}^{q} b_l u(n-l) \tag{3.1}$$

where G is the gain factor and a_k and b_k are the so-called filter coefficients (where $b_0 = 1$ is assumed). Applying the z-transform to the above equation, the transfer function of the vocal tract, can be obtained as follows:

$$H(z) = \frac{X(z)}{U(z)} = G \frac{1 + \sum_{l=1}^{q} b_l z^{-1}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.2}$$

where we define the z-transform of signal $x(n)$ as $X(z) = \sum_{n=-\infty}^{n=\infty} x(n) z^{-n}$ and similarly for $U(z)$ in terms of $u(n)$. As it can be seen, the transfer function corresponds to a pole-zero model, where in the context of the human auditory system, the zeros represent the nasals and the poles represent the formants in a vowel spectrum. In order to reduce the complexity, it is generally assumed that the transfer function has no zeros. This is referred to an all-pole or AR model. Therefore, the all-pole transfer function can be represented:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{3.3}$$

Or equivalently, the time-domain equation is derived as:

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + Gu(n \tag{3.4}$$

where filter coefficients a_k are called LP or AR coefficients. The error signal $e(n)$, also called residual error, is defined as the difference between the output signal and its predicted value, that is:

$$e(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k) = Gu(n) \tag{3.5}$$

In z-domain (3.5) is equivalent to:

$$E(z) = X(z)A(z) \tag{3.6}$$

where we define $A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$.

On the above basis, it can be concluded that LP provides an analysis-synthesis framework for speech signals. The analysis system takes the speech signal $x(n)$ as the input to a spectral shaping filter, with transfer function $H(z)$, in order to produce the error signal $e(n)$. Alternatively, the error signal $e(n)$ can be fed to the synthesis system, where the input is filtered by $1/H(z)$ to produce the speech signal $x(n)$. These operations are demonstrated in Figure 3.2 [48].

In practice, the order p and coefficients $a_k, k \in \{1..p\}$ of the LP analysis system should be chosen such that the residual error $e(n)$ has the character of white noise, that is $e(n)$ should have a zero-mean and an impulse like correlation function, as in $E[e(n)e(m)] = \sigma^2\delta(n-m)$, or that the resulting PSD level is constant ad equal to σ^2 for all frequencies, Under this condition an accurate representation of the speech signal $x(n)$ PSD say $P_x(\omega)$ can be achieved using suitable LP coefficients $a_k, k \in \{1..p\}$ and residual error power σ^2. In practice, it can be shown that

$$P_x(\omega) = \frac{\sigma_x^2}{|A_x(e^{(j\omega)})|^2} \tag{3.7}$$

Therefore, an accurate representation of the speech signal spectrum can be achieved using suitable LP coefficients and residual error variance.

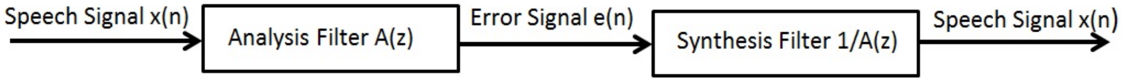


Fig. 3.2 LP analysis and synthesis model

3.1.2 Codebook generation using Generalized Lloyd vector quantization method

As explained before, trained codebooks of speech and noise LP coefficients can be used to model the *a priori* information needed in the process of speech enhancement. In [49] these codebooks are generated using a vector quantization (VQ) method called Generalized Lloyd algorithm (GLA).

In general, a vector quantizer of dimension K and size S is described as a mapping from the K dimensional data vectors defined in Euclidean space to a finite subset C, in which C includes N vectors called codevectors. The set C of all codevectors is called the codebook, while N is the size of the codebook. The GLA method is one of the most popular VQ algorithms for codebook generation. It was first introduced in [49] and is also named as LBG based on the initials of the authors of the paper Linde, Buzo and Gray.

LBG is an iterative clustering algorithm, which produces an optimum codebook for a given data source, by minimizing a distortion measure between a training vector and the codevector which is closest to it.

In order to apply this algorithm, there is a need for a training sequence which provides the basic model for the data to be encoded. The training sequence is usually obtained from a large database, for example in the case of the speech signals, the training sequence can be obtained by recording several long conversations. When applying the LBG algorithm, the training sequence is partitioned into several groups based on the codebook size. Among all the vectors in a specific group, a centroid vector is chosen to be the codevector, based on minimization of a distortion measure. The centroid vector is actually the representative of that group. The main goal of LBG is to minimize the distortion measure between the training vectors and their representation codevectors, that is: finding the optimal partition and codevectors to minimize the overall distortion.

In effect, LBG employs and iterative procedure which is repeated until the averaged distortion is minimized. The main step of this process can be summarized as follows [50]:

1) An initial codebook containing N codevectors is first chosen.

2) The training sequence is partitioned into N groups using the distortion measure.

3) For each group, a centroid vector is selected to get an improved codevector.

4) Step 2 is repeated if the distortion measure is larger than the minimum average considered as the threshold.

Different methods have been proposed in order to generate the initial codebook in step 1. The original method applied in LBG is called binary splitting. In this method, an initial codevector is obtained as the average of the entire training sequence. This codevector is then split into two and the iterative algorithm is run with these two vectors as initial codebooks. The final two code vectors are split into four and the process is repeated until the desired number of codevectors is obtained.

In order to implement the codebook based method for estimating speech and noise statistics, the GLA method will be applied. For this application, the training sequence will include vectors of LP coefficients obtained from different male and female recorded speech segments, and then different codebooks of speech and noise LP coefficients are produced. The methodology used for codebook generation will be discussed in further Chapter 4.

3.1.3 Codebook based ML parameter estimation

In this section, we discuss the estimation of excitation variances of the speech and noise AR models, based on the *a priori* information stored in the codebooks. At first, it is assumed that the excitation variances are unknown deterministic parameters. Consequently, an ML based method is used to estimate the excitation variances of the speech and noise.

Assuming that we have an additive noise model as in (2.1), the general idea of the ML-based parameter estimation method is to search through the codebooks in order to derive the maximum likelihood estimates of speech and noise excitation variances. The main step of this procedure can be summarized as follows [42]:

- For each pair of the speech and noise LP coefficients in the codebook, the speech and noise excitation variances which maximize the likelihood function are derived.

- Using the derived excitation variances and associated LP coefficients, the AR spectra are constructed using the AR model as explained in Section 3.1.1.

- Based on these AR spectra, a log-likelihood score which represents the error between the modeled spectra and the measured powers in the given time frame, is computed.

- A search is done through all the computed log-likelihood scores in order to find the spectra which maximize the score.
- The excitation variances associated with these spectra, are considered as maximum likelihood estimates of the speech and noise excitation variances.

Figure 3.3 provides a schematic diagram of the above method, which is explained in further mathematical detail below.

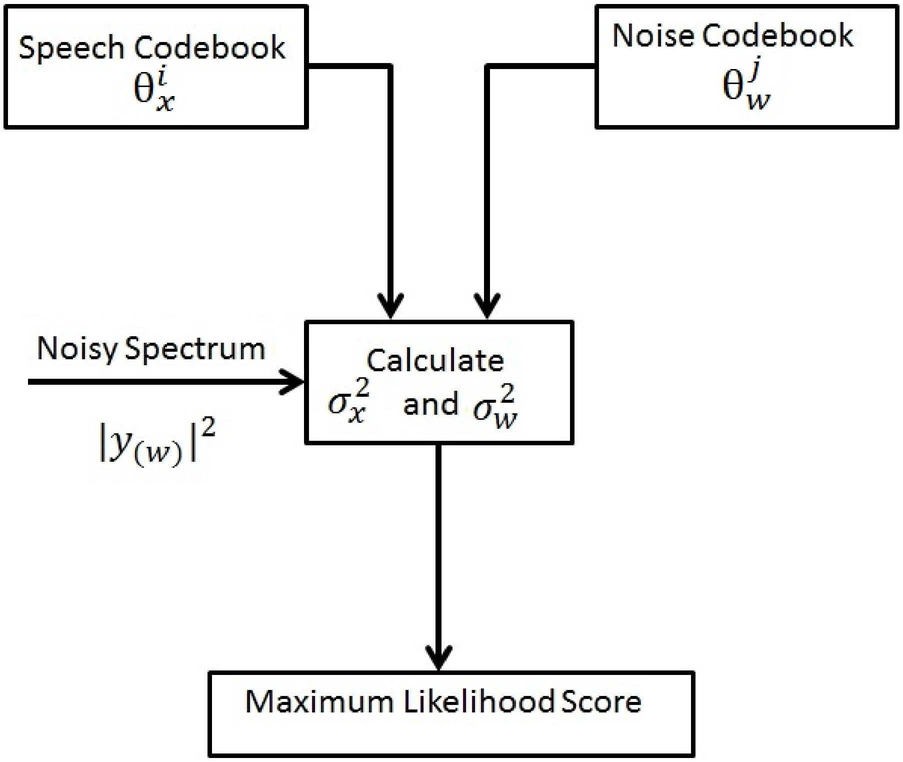

Speech Codebook θ_x^i
Noise Codebook θ_w^j
Noisy Spectrum
$|y_{(w)}|^2$
Calculate σ_x^2 and σ_w^2
Maximum Likelihood Score

Fig. 3.3 ML scheme

Let θ_x^i and θ_w^j represent the ith and jth codebook vectors of speech and noise LP coefficients, respectively which are defined as:

$$\boldsymbol{\theta}_x^i = (a_{x_0}^i, \cdots, a_{x_p}^i), \quad \boldsymbol{\theta}_w^j = (a_{w_0}^j, \cdots, a_{w_q}^j) \tag{3.8}$$

where p and q are the LP orders of the speech and noise AR models, respectively and $a_{x_k}^i$ and $a_{w_k}^j$ are the corresponding LP coefficients. The ML estimates of the speech and noise excitation variances are obtained according to :

$$\{\sigma_x^{2*}, \sigma_w^{2*}\} = \arg \max_{i,j,\sigma_x^2,\sigma_w^2} p_y(\boldsymbol{y}|\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j; \sigma_x^2, \sigma_w^2) \tag{3.9}$$

where $\boldsymbol{y} = [y(0), y(1)...y(L-1)]^T$ is the vector of observed noisy speech samples in the given frame and L is the frame length.

Assuming that the conditional PDF in (3.9) is Gaussian, the likelihood function for each frame of the noisy speech can be written as [42]:

$$p_y(\boldsymbol{y}|\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j; \sigma_x^2, \sigma_w^2) = \frac{1}{(2\pi)^{L/2}|R_y|^{1/2}} e^{-(1/2)(\boldsymbol{y}^t R_y^{-1} \boldsymbol{y})} \tag{3.10}$$

where R_y is the noisy speech covariance matrix which is defined as the sum of the speech and noise covariance matrices $R_y = R_x + R_w$.

In the following paragraph, we summarize the developments from [42] leading to an approximate ML solution for the optimal gains σ_x^2 and σ_w^2 corresponding to a particular codebook pair (i, j).

Using (3.10) the log- likelihood function (LLF) can be written as:

$$l(\sigma_x^2, \sigma_w^2) = \ln P_y(\boldsymbol{y}|\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j; \sigma_x^2, \sigma_w^2) = C - \frac{1}{2} ln|R_y| - \frac{1}{2} y^T R_y^{-1} \boldsymbol{y} \tag{3.11}$$

where $|R_y|$ denotes the determinant of R_y.

It is possible to simplify the LLF (3.11), using properties of the Toeplitz Hermitian matrices developed in [51]. Therefore a simplified version of the LLF only dependent on the excitation variances and normalized spectra can be written as [42]:

$$l(\sigma_x^2, \sigma_w^2) = \int_0^{2\pi} \left(\left(\frac{\sigma_x^2 |A_x^i|^2 + \sigma_w^2 |A_w^j|^2}{|A_x^i|^2 + |A_w^j|^2} \right) + P_y \left(\frac{|A_x^i|^2 |A_w^j|^2}{\sigma_x^2 |A_x^i|^2 + \sigma_w^2 |A_w^j|^2} |A_x^i|^2 + |A_w^j|^2 \right) \right) d\omega \tag{3.12}$$

where A_x^i and A_w^j are the ith speech codebook and jth noise codebook spectra, previously defined as:

$$A_x^i \equiv A_x^i(\omega) = \sum_{k=0}^{p} a_{x_k}^i e^{-j\omega k}, \quad A_w^j \equiv A_w^j(\omega) = \sum_{k=0}^{p} a_{w_k}^j e^{-j\omega k} \tag{3.13}$$

where $a^i_{x_k}$ and $a^j_{w_k}$ represent the LP coefficients of the speech and noise, respectively and $P_y \equiv P_y(\omega) = |Y(\omega)|^2$ with $Y(\omega) = \sum_{k=0}^{L-1} y[k]e^{(-j\omega k)}$. As we are looking for the gains which make the LLF maximum, we should set the partial derivatives of (3.12) to zero which yields:

$$\int_0^{2\pi} \frac{|A_w|^2(P_y|A_x|^2|A_w|^2 - \sigma_x^2|A_w|^2 - \sigma_w^2|A_x|^2)}{\sigma_x^2|A_w|^2 + \sigma_w^2|A_x|^2} d\omega = 0 \tag{3.14}$$

$$\int_0^{2\pi} \frac{|A_x|^2(P_y|A_w|^2|A_x|^2 - \sigma_x^2|A_w|^2 - \sigma_w^2|A_x|^2)}{\sigma_x^2|A_w|^2 + \sigma_w^2|A_x|^2} d\omega = 0 \tag{3.15}$$

These equations can be solved exactly; only if the codebooks contain spectral shapes A^i_x and A^j_w that fulfill the following condition:

$$P_y = \frac{\sigma_x^2}{|A_x|^2} + \frac{\sigma_w^2}{|A_w|^2}, \qquad \forall \in [0, 2\pi] \tag{3.16}$$

To obtain a solution in practice, let us assume that (3.16) is fulfilled with small error $\sigma = Z - Z_0$, where we define:

$$Z_0 = P_y|A^i_x|^2|A_w j|^2 \tag{3.17}$$

$$Z = \sigma_x^2|A^j_w|^2 + \sigma_w^2|A^i_x|^2 \tag{3.18}$$

Then we can see how the LLF in (3.12) behaves in the neighborhood of the maximum. Substituting (3.18) into (3.12), and using a Taylor series expansion of (3.12) around $\sigma = 0$, it can be shown that [42]:

$$l(\sigma) = \int_0^{2\pi} (1 + \ln(P_y) + \frac{1}{2(P_y|A_x|^2|A_w|^2)^2}\sigma^2) d\omega + O(\sigma^3) \tag{3.19}$$

Hence, for small σ, the LLF depends only on the weighted squared error l^2. Assuming that the effect of the weight $P_y^2|A^i_x|^4|A^j_w|^4$ can be neglected, maximization of (3.19) is equivalent to:

$$\arg\min_{\sigma_x^2, \sigma_w^2} (||P_y|A_x|^2|A_w|^2 - \sigma_x^2|A_w|^2 - \sigma_w^2|A_x|^2||^2) \tag{3.20}$$

where we define $\|f(\omega)\|^h = \int_0^{2\pi} |f(\omega)|^h d\omega$ for $h > 0$. Setting the partial derivatives with respect to σ_x and σ_w to zero, the excitation variances are obtained as the solution to the following linear system of equations:

$$C \begin{bmatrix} \sigma_x^2 \\ \sigma_w^2 \end{bmatrix} = D \tag{3.21}$$

$$C = \begin{bmatrix} \left\| |A_w^j(\omega)|^4 \right\| & \left\| |A_x^i(\omega)|^2 |A_w^j(\omega)|^2 \right\| \\ \left\| |A_x^i(\omega)|^2 |A_w^j(\omega)|^2 \right\| & \left\| |A_x^i(\omega)|^4 \right\| \end{bmatrix} \tag{3.22}$$

$$D = \begin{bmatrix} \left\| P_y(\omega) |A_x^i(\omega)|^2 |A_w^j(\omega)|^4 \right\| \\ \left\| P_y(\omega) |A_x^i(\omega)|^4 |A_w^j(\omega)|^2 \right\| \end{bmatrix} \tag{3.23}$$

Having found the excitation variances corresponding to a given pair of speech and noise spectra in the codebook, that is, A_x^i and A_w^j, the LLF score based on (3.12) can be evaluated. Consequently, the combination of the speech and noise spectra, that is the pair of codebook indexes i and j which maximize the likelihood score is considered as the maximum likelihood estimates of the speech and noise spectra.

3.1.4 MMSE estimation of short time predictive (STP) parameters

In the previous section, the excitation variances and AR parameters were treated as unknown deterministic parameters. A possible alternative is to consider the gains and AR parameters to be random variables, which should be estimated based on the characteristics of their PDF. In this regard, the ML estimates derived in Section 3.1.3 can be exploited in defining the *a priori* distributions of the speech and noise parameters needed in the Bayesian framework. Below we summarize this technique which was originally exposed in [44].

In this formulation, the parameters subject to estimation are the LP coefficients of the speech and noise, as represented by vectors

$$\boldsymbol{\theta_x} = [a_{x_0}, ..., a_{x_p}], \qquad \boldsymbol{\theta_w} = [a_{w_0}, ..., a_{w_q}] \tag{3.24}$$

and respectively the associated excitation variances σ_x^2 and σ_w^2. This complete set of parameters can be represented by a single vector

$$\boldsymbol{\theta} = [\boldsymbol{\theta_x}, \boldsymbol{\theta_w}, \sigma_x^2, \sigma_w^2] \tag{3.25}$$

which is modeled as a random entity with joint PDF $p(\boldsymbol{\theta})$. The main goal here is to estimate $\boldsymbol{\theta}$ on the observed noisy speech samples contained in vector $\boldsymbol{y}$.

In the MMSE approach, we seek an estimator of $\boldsymbol{\theta}$ which minimizes the mean square error $E\{\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}|^2\}$. The solution to this problem is given by the conditional expectation [52].

$$\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}(\boldsymbol{y}) = E\{\boldsymbol{\theta}|\boldsymbol{y}\} \tag{3.26}$$

Expanding the expected value, we can rewrite (3.26) as follows:

$$\hat{\boldsymbol{\theta}} = \int_{\Theta} \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} = \int_{\Theta} \boldsymbol{\theta} \frac{p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{y})} d\boldsymbol{\theta} \tag{3.27}$$

where Θ is the support-space of the LP coefficients and excitation variances, defined as $\Theta = \Theta_x \times \Theta_w \times \sum_x \times \sum_w$ where Θ_x and Θ_w represent the support-space of the vectors of the LP coefficients of speech and noise, and $\sum_x$ and $\sum_w$ are the support-space for the speech and noise excitation variances. [44].

Based on the definition of (3.10), the conditional probability $p(\boldsymbol{y}|\boldsymbol{\theta})$ in (3.27) can be modeled as a zero-mean Gaussian with covariance matrix $R_x + R_w$. The speech covariance matrix R_x can be defined as $R_x = \sigma_x^2 (A_x^T A_x)^{-1}$, where A_x is the $L \times L$ lower triangular Teoplitz matrix in which the first column is $[1, a_{x_1}, a_{x_2} \cdots a_{x_p}, 0 \cdots 0]$ and L is the frame length. The noise covariance matrix R_w can be defined similarly.

As explained before, we assume that the speech and noise samples are statistically independent. Therefore, the PDF of speech and noise STP parameters are also assumed to be independent.

This can be shown as:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_x, \sigma_x^2)p(\boldsymbol{\theta}_w, \sigma_w^2). \tag{3.28}$$

It can also be assumed that the spectral shapes and gains are independent, therefore $p(\boldsymbol{\theta}_x, \sigma_x^2) = p(\boldsymbol{\theta}_x)p(\sigma_x^2)$ and the same for the noise. It is proved in [44] that the conditional probability $p(\boldsymbol{y}|\boldsymbol{\theta})$ decays rapidly from its maximum value as a function of the deviation from the true excitation variances, which are approximated by the ML estimates denoted by $\sigma_x^{2,ML}$ and $\sigma_w^{2,ML}$, and derived in the previous section. Therefore the conditional probability $p(\boldsymbol{y}|\boldsymbol{\theta})$ is approximated by $p(\boldsymbol{y}|\boldsymbol{\theta})\delta(\sigma_x^2 - \sigma_x^{2,ML})\delta(\sigma_w^2 - \sigma_w^{2,ML})$, where $\delta(.)$ is the Dirac-delta function. Therefore, the equation (3.27) can be approximated as follows [44]:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &\approx \int_{\Theta} \boldsymbol{\theta} \frac{p(\boldsymbol{y}|\boldsymbol{\theta})\delta(\sigma_x^2 - \sigma_x^{2,ML})\delta(\sigma_w^2 - \sigma_w^{2,ML})p(\boldsymbol{\theta}_x)p(\boldsymbol{\theta}_w)}{p(\boldsymbol{y})} d\boldsymbol{\theta} \\
&= \int_{\Theta_x}\int_{\Theta_w} \boldsymbol{\theta} \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_x, \boldsymbol{\theta}_w, \sigma_x^{2,ML}, \sigma_w^{2,ML})p(\boldsymbol{\theta}_x)p(\boldsymbol{\theta}_w)}{p(\boldsymbol{y})} d\boldsymbol{\theta_x} x d\boldsymbol{\theta_w} \qquad (3.29)
\end{aligned}$$

It should be noticed that the support-space in (3.29) has been reduced to the support-space of the two of LP vectors. Using the approximated conditional probability $p(\boldsymbol{y}|\boldsymbol{\theta})$, the PDF of $\boldsymbol{y}$ can be obtained as follows:

$$p(\boldsymbol{y}) = \int_{\Theta_x}\int_{\Theta_w} p(\boldsymbol{y}|\boldsymbol{\theta}_x, \boldsymbol{\theta}_w, \sigma_x^{2,ML}, \sigma_w^{2,ML})p(\boldsymbol{\theta}_x)p(\boldsymbol{\theta}_w)d\theta_x d\theta_w \tag{3.30}$$

Using numerical integration, (3.29) and (3.30) can be evaluated with the help of the trained codebooks as follows:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \frac{1}{N_x N_w} \sum_{i,j=1}^{N_x,N_w} \boldsymbol{\theta'}_{i,j} \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j, \sigma_{x,ij}^{2,ML}, \sigma_{w,ij}^{2,ML})p(\boldsymbol{\theta}_x^i)p(\boldsymbol{\theta}_w^j)}{p(\boldsymbol{y})} \qquad (3.31) \\
p(\boldsymbol{y}) &= \frac{1}{N_x N_w} \sum_{i,j=1}^{N_x,N_w} p(\boldsymbol{y}|\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j, \sigma_{x,ij}^{2,ML}, \sigma_{w,ij}^{2,ML})p(\boldsymbol{\theta}_x^i)p(\boldsymbol{y}|_w^j) \qquad (3.32)
\end{aligned}$$

where $\boldsymbol{\theta'}_{i,j}$ is defined as $\boldsymbol{\theta'}_{i,j} = [\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j, \sigma_{x,ij}^{2,ML}, \sigma_{w,ij}^{2,ML}]$, $\boldsymbol{\theta}_x^i$ and $\boldsymbol{\theta}_w^j$ are the ith speech codebook and $j$the noise codebook, $\sigma_{x,ij}^{2,ML}$ and $\sigma_{w,ij}^{2,ML}$ are the ML estimates of speech and noise variances based on the ith speech codebook entry and jth noise codebook entry as given by (3.21), and N_x and N_w are the speech and noise codebook sizes.The quantities $p(\boldsymbol{\theta}_x^i)$ and $p(\boldsymbol{y}|_w^j)$ denote the *a priori* PDF of the speech and noise AR parameters, evaluated at the codebook entries. These can be approximated as $\frac{N_x^i}{N_x}$ and $\frac{N_w^j}{N_w}$, where N_x^i and N_w^j are the number of training vectors in respective Voronoi cells. These probabilities are constant if the number of training vectors in each Voronoi cells is the same [53], assumed to be the case in this thesis.

Along with the log-likelihood measure, another well-known distortion measure is the Itakura-Saito distance [54]. It is shown in [54] that maximizing the log-likelihood, is equivalent to minimizing the Itakura-Saito distance between the spectra of the observed noisy data $\boldsymbol{y}$ and the modeled spectra. The modeled spectra can be defined as follows:

$$\hat{P}_y = \frac{\sigma_x^2}{|A_x(\omega)|^2} + \frac{\sigma_w^2}{|A_w(\omega)|^2} \tag{3.33}$$

where A_x^i and A_w^j, defined in (3.13) are the AR spectra of the speech and noise. as follows:

$$A_x^i = \sum_{k=0}^{p} a_{x_k}^i e^{-j\omega k} \tag{3.34}$$

$$A_w^j = \sum_{k=0}^{q} a_{w_k}^j e^{-j\omega k} \tag{3.35}$$

The Itakura-Saito measure between the two spectra is defined as follows [54]:

$$d_{IS}(P_y, \hat{P}_y) = \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{P_y(\omega)}{\hat{P}_y(\omega)} - \ln(\frac{P_y(\omega)}{\hat{P}_y(\omega)}) - 1\right)d\omega \tag{3.36}$$

Based on the above assertion, it is shown in [44] that the conditional PDF in (3.32) can be expressed as follows:

$$p(\boldsymbol{y}|\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^i, \sigma_{x,ij}^{2,ML}, \sigma_{w,ij}^{2,ML}) = C\exp(-d_{IS}(p_y, \hat{p}_y^{i,j,ML}) \tag{3.37}$$

where C is a numerical constant that will cancel out from the nominator and denominator in(3.32) so that is the value is not important. The codebook combination which produces negative values for the speech or noise excitation variances should be neglected due to the nonnegativity constraints on the variances [43].

Having computed $\hat{\theta}$ based on (3.32) it is possible to build the modeled spectra of speech and noise according to AR model in (3.33). Once the desired speech and noise LP power spectra has been completed, they are fed into a speech enhancement system, which evaluates the estimated clean speech spectrum. This is further discussed in Section 3.2.

3.2 Incorporation of the Codebook Based STP Parameter Estimation into the Wβ-SA Method

As discussed in Section 2.2, in the Bayesian STSA methods of speech enhancement, a gain function is applied to the spectral magnitude of the noisy speech in order to find the estimate of the clean speech spectrum. The Bayesian speech enhancement used in this thesis is the Wβ-SA method from [39] as described in Section 2.2.2. The gain function utilized in this method is given by the following expression:

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \left(\frac{\Gamma\left(\frac{\beta}{2} - \alpha + 1\right) M\left(\alpha - \frac{\beta}{2}, 1; -\upsilon_k\right)}{\Gamma\left(-\alpha + 1\right) M\left(\alpha, 1; -\upsilon_k\right)} \right)^{1/\beta} \tag{3.38}$$

where the physical interpretation of the various parameters, i.e. α, β, γ and ξ, has been previously given in Section 2.2.

As it can be observed from (3.38), the gain G_k is a function of γ_k and υ_k, where in turn υ_k is a function of the *a posteriori* and *a priroi* SNRs γ and ξ as defined in (2.51). Therefore, it can be concluded that only the two values of γ_k and ξ_k need to be evaluated to compute the clean speech estimate. The values of these two parameters are defined in (2.51) and reproduced here for convenience:

$$\xi_k = \frac{\sigma^2_{X,k}}{\sigma^2_{W,k}}, \qquad \gamma_k = \frac{|Y_k|^2}{\sigma^2_{W,k}}. \tag{3.39}$$

As it can be seen these two parameters are dependent on the noise and speech variances denoted by $\sigma^2_{X,k}$ and $\sigma^2_{W,k}$.

Equation (3.39) provides the required link for incorporating the codebook based STP parameter estimation into the Wβ-SA speech enhancement method. Specifically, the variances of the speech and noise can be approximated by the speech and noise PSD, derived in Section 3.1.4.

Therefore recalling our discussion on the speech and noise PSD from Section 3.1.4, we obtain the following expressions:

$$\sigma^2_{X,k} = \hat{p}_x(\omega_k) = \frac{\sigma^2_{x,k}}{|A_x(\omega_k)|^2} \qquad \sigma^2_{W,k} = \hat{p}_w(\omega_k) = \frac{\sigma^2_{w,k}}{|A_w(\omega_k)|^2} \tag{3.40}$$

Using (3.40), it is possible to apply the speech and noise PSDs estimated from the codebook based approach presented in Section 3.1 to calculate the gain function (3.38), which is needed for the application of the Wβ-SA method on each frame. The complete procedure is summarized in the following block diagram.

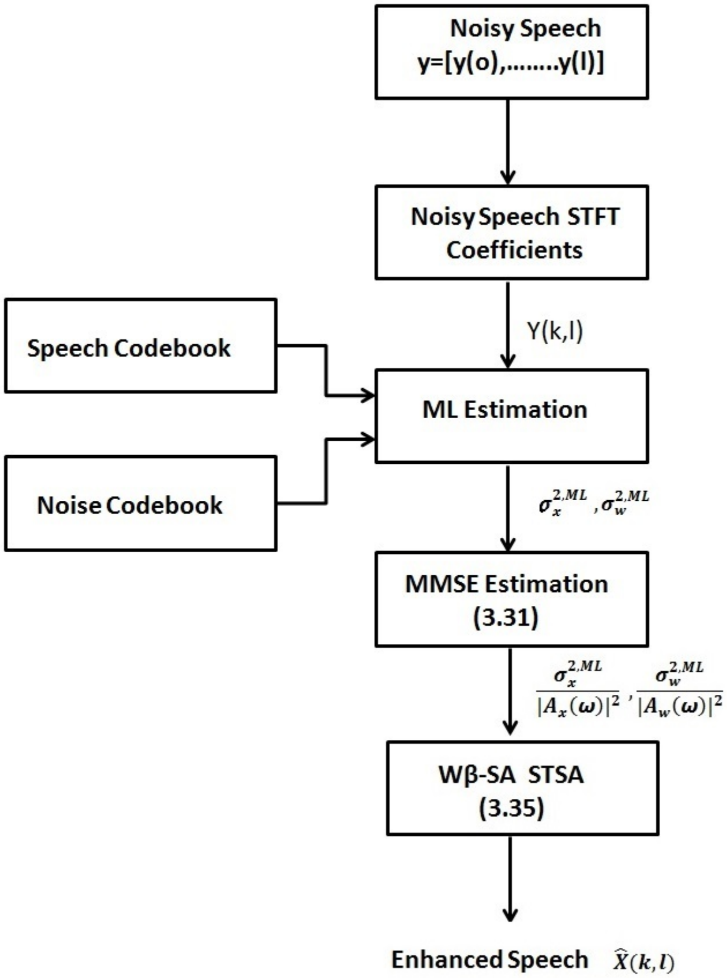

Noisy Speech
y=[y(o),........y(l)]
Noisy Speech STFT Coefficients
Y(k,l)
Speech Codebook
Noise Codebook
ML Estimation
$\sigma_x^{2,ML}$, $\sigma_w^{2,ML}$
MMSE Estimation (3.31)
$\frac{\sigma_x^{2,ML}}{|A_x(\omega)|^2}, \frac{\sigma_w^{2,ML}}{|A_w(\omega)|^2}$
Wβ-SA STSA (3.35)
Enhanced Speech $\hat{X}(k,l)$

Fig. 3.4 Block diagram of the complete procedure for Wβ-SA speech enhancement using codebook based STP estimation

3.2.1 Decision-directed estimation approach

A major problem that arises in the application of various speech enhancement methods is the production of musical noise. Musical noise is a perceptual phenomenon characterized by tones at different frequencies, which appear and disappear haphazardly, and can be extremely annoying to a human listener.This occurs when the enhancement algorithm is too aggressive in removing the noise, which tends to the production of musical noise [39]. Some methods have been introduced in the literature to reduce the musical noise produced in Bayesian speech enhancement methods. Some of these methods, focus on computing the *a priori* SNR, or ξ_k, using some alternative approaches instead of its definition (3.39). The method we applied in this thesis is called decision-directed and was first introduced in [30]; it is explained in further detail below.

On the other hand, as discussed before, the *a priori* SNR ξ is given by its definition in (3.39) as the ratio of the clean speech variance to that of the noise. On the other hand, under the independence assumption for the speech signal and noise, the relationship between the *a priori* and *a posteriori* SNR can be expressed as:

$$\xi_k = E\{\gamma_k - 1\}. \tag{3.41}$$

Combining the two equations, we can obtain a recursive estimator of ξ_k at the ℓ-th frame, denoted by $\xi(k, l)$, via the following operation:

$$\xi(k, \ell) = \tau \frac{G(k, \ell - 1)^2 |Y(k, \ell - 1)|}{\sigma_W^2(k, \ell - 1)} + (1 - \tau) \max[\gamma(k, \ell) - 1, 0] \tag{3.42}$$

where τ is a weighting or smoothing factor in the range of $0.95 \leq \tau < 1$. The max[.,.] operation prevents negativity in the instantaneous SNR, i.e. $\gamma(k, \ell) - 1$.

In this work we used a slightly modified version of (3.42) as follows:

$$\xi(k, \ell) = \tau \frac{G(k, \ell - 1)^\rho |Y(k, \ell - 1)|}{\sigma_W^2(k, \ell - 1)} + (1 - \tau) \frac{\sigma_X^2(k, \ell)}{\sigma_W^2(k, \ell)} \tag{3.43}$$

where τ is a weighting or smoothing factor in the range of $0.95 \leq \gamma < 1$, and $\sigma_X^2(k, \ell)$ and $\sigma_W^2(k, \ell)$ are the speech and noise PSDs obtained from the codebook approach at the ℓth frame. In practice, we found that a value of ρ smaller than one produces better results. This nonlinear smoothing has the great advantage of eliminating large power variations in consecutive frames, which in turn tend to reduce the level of musical noise [55]. The processed speech will be more comfortable for a listener while listening to the enhanced speech file.

Chapter 4

Experimental Results

In this chapter, we investigate the performance of the newly proposed speech enhancement method developed in Chapter 3 by presenting the results of selected experiments, including both objective and subjective performance evaluations. In all of the experiments, we compare the new method, which combines the Wβ-SA speech enhancement with the codebook-based approach for estimating the noise and speech statistics, with a similar but alternative scheme where the Wβ-SA is replaced by the Wiener filter [29]. We begin by describing the methodology and then present and discuss the results so obtained.

4.1 Methodology

An 8-bit speech codebook of LP coefficients of dimension p=10 was trained using the generalized Lloyd algorithm (GLA) as explained in Section 3.1.2. The speech training set consists of 4 minutes of recorded clean speech from 2 male and 2 female speakers, available from the McGill TSP database [56]. A 3-bit noise codebook of LP coefficients of dimension q=10 was trained in the same way. It may be argued that the use of 3 bits for the noise codebook size is small. However, this value is consistent with those used in [43] where codebook of sizes 1 up to 4 bits (i.e. 2 up to 16 entries) were found to be optimal for different types of noise, e.g. highway, white, babble, and siren noise. The noise training set consists of a 10 minutes concatenation of five different types of recorded noise from the AURORA database [57], that is: car, train, street, restaurant and airport noise. Some of the noise files in the training set may contain background speech. For example the train noise file contains a brief announcement while the restaurant noise file might contain some low level of barely discernible voice sounds from people speaking. We understand that the presence of speech in the noise samples used in the training set might impair the quality of the noise codebook, but we could not further investigate this aspect, which remains open for future work.

Selected noise segments were added to the clean speech and enhancement experiments were conducted for noisy speech with input SNR of 0, 5, 10 and 15 dB. In this thesis, we used a simple measure of SNR, defined as the ratio of total speech energy to noise energy. However, in speech processing applications, since the speech may have gaps or periods of silence, more sophisticated measures of SNR can be used that only consider non-silent portions of the speech signals [58].

The sampling frequency of the speech and noise signals is set to F_s=8kHz. In the application of the STFT processing, a frame length of N=256 samples with 50% overlap is used, where the frames are windowed using a Hanning window. In the STFT domain, the enhancement of the noisy speech is carried out by applying a gain to the noisy speech STFT, as (2.49). We compare two different methods namely:

- Combination of Wβ-SA with codebook based estimation of spectral statistics, as described in Section 3.1.4.
- Combination of Wiener filter [29] with codebook based approach where the enhancement gain is given by:

$$G_k = \frac{\sigma^2_{x,k}}{|A_x(\omega_k)|^2} / (\frac{\sigma^2_{x,k}}{|A_x(\omega_k)|^2} + \frac{\sigma^2_{w,k}}{|A_w(\omega_k)|^2}) \tag{4.1}$$

The two algorithms are first compared in terms of a well-known objective measure, i.e. perceptual evaluation of speech quality (PESQ). This measure is recommended by ITU-T for speech quality [59], and it is generally well correlated with subjective results.

In the process of implementing the proposed method, we found that the noise PSD obtained by the codebook based algorithm tends to be underestimated and that better results could be obtained by applying a multiplicative bias, whose value was found and adjusted through separate experiments. Regarding (3.43), the best results were obtained by setting $\tau = 0.85$ and $\rho = 1$. In this approach, once the codebooks have been trained properly, it is possible to estimate the noise PSD without performing any specific estimation of noise parameters on a noise-only preamble to the noisy speech signal.

In Section 4.2, first the accuracy of the speech and noise codebook spectra is examined by comparing them with the actual power spectra of the speech and noise. It will be observed that the codebook spectra lie in the same range as the actual ones. The MMSE estimates of the noise spectrum, based on the codebook approach in Section 3.1.4, are then compared with the actual noise power spectrum for different scenarios. It will be seen that the algorithm performs well in estimating the noise power spectrum. Afterwards, the waveforms of the noisy, clean and enhanced speech are compared in order to evaluate the algorithm performance in removing background noise. It will be observed that better results are obtained with the proposed approach compared to Wiener filter. Finally, in Section 4.4, the results of the two methods are compared subjectively, where the listeners mostly prefer the enhancement speech obtained with our proposed approach.

4.2 Numerical Experiments

4.2.1 Accuracy of the trained codebooks

At first, we need to ensure that the speech and noise codebooks are generated properly. In order to demonstrate the validity of the codebooks, we examine whether the spectra from the speech and noise codebook entries fall in the same range as the real spectra of the speech and noise.

In Figure 4.1, the LP spectrum from a selected portion of train noise and the LP spectra obtained from the corresponding noise codebook entries are depicted. As it can be observed, the noise codebook is generated such that the spectra associated with its entries follow the same pattern as the real spectrum of the noise, which is a desirable property.

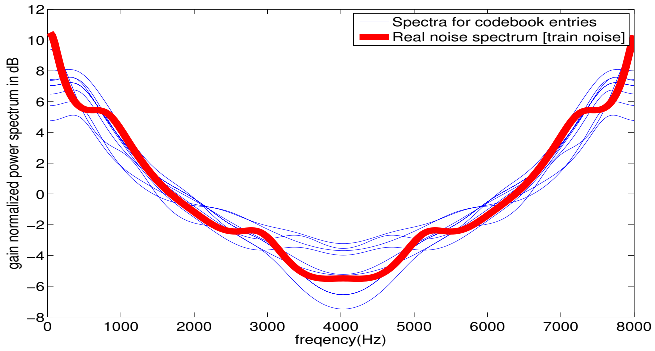


Fig. 4.1 Plot of the true noise LP power spectrum and the noise codebook entries LP spectra

In Figure 4.2 the same calculations are repeated for the speech codebook. This figure shows the LP spectrum of a selected speech utterance from one of the female speakers whose recorded speech is included in the training set, along with the LP spectra obtained from the speech codebook. The spectrum of the speech codebook entry which is closest (in the squared error sense) to the spectrum of the true speech is also shown in 4.2. The spectra of the speech codebook entries fall in the same range as the true speech spectrum, which is again a desirable property.

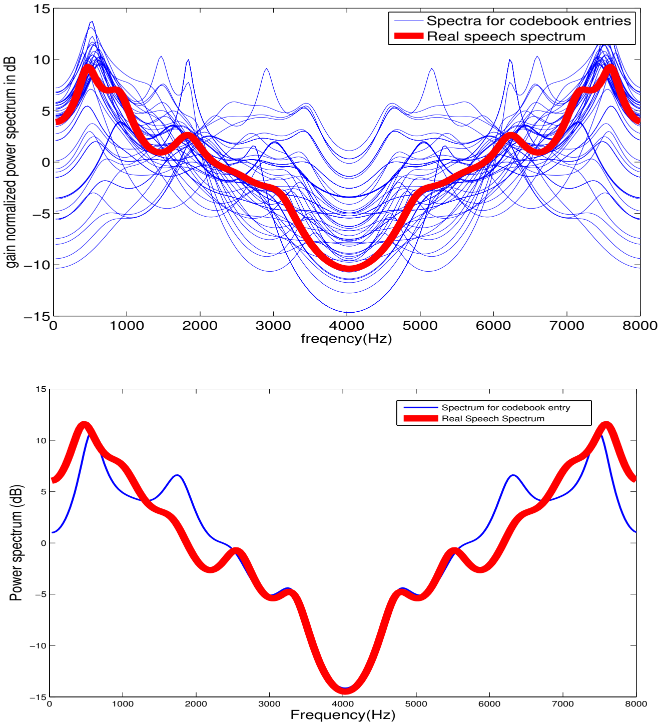


Fig. 4.2 Plot of the true speech LP power spectrum and the speech codebook entries spectra. Top: all the codebook entries; Bottom: the best match between speech spectrum and speech codebook entry spectrum

4.2.2 Accuracy of the noise estimation

As it has been discussed before, a crucial component of any speech enhancement system is the estimation of the noise statistics. In this section, the quality of the codebook based noise spectrum estimation is examined by comparing the complete codebook based estimator, as given by $\sigma^2_{W,k}$ in (3.40) which incorporates both the LP modeling and the gain estimator, to the STFT-based periodogram derived from the corresponding noise-only data frame. The results are given for different scenarios, i.e. different speakers, noise types and SNR.

In Figure 4.3, the estimated noise spectrum from a noisy speech is compared with the actual STFT-based noise power spectrum. Four different types of noise are considered, namely: train, car, street and airport noise. In all four cases, the noise is added to the clean speech from a female speaker with SNR=0dB to obtain the noisy speech file. The figure shows the noise power spectrum estimated with the codebook based approach along with the periodogram based spectrum for a selected frame of the noisy speech. It can be observed from these various plots that the codebook based algorithm can estimate relatively well the spectral envelopes of these various noise types.

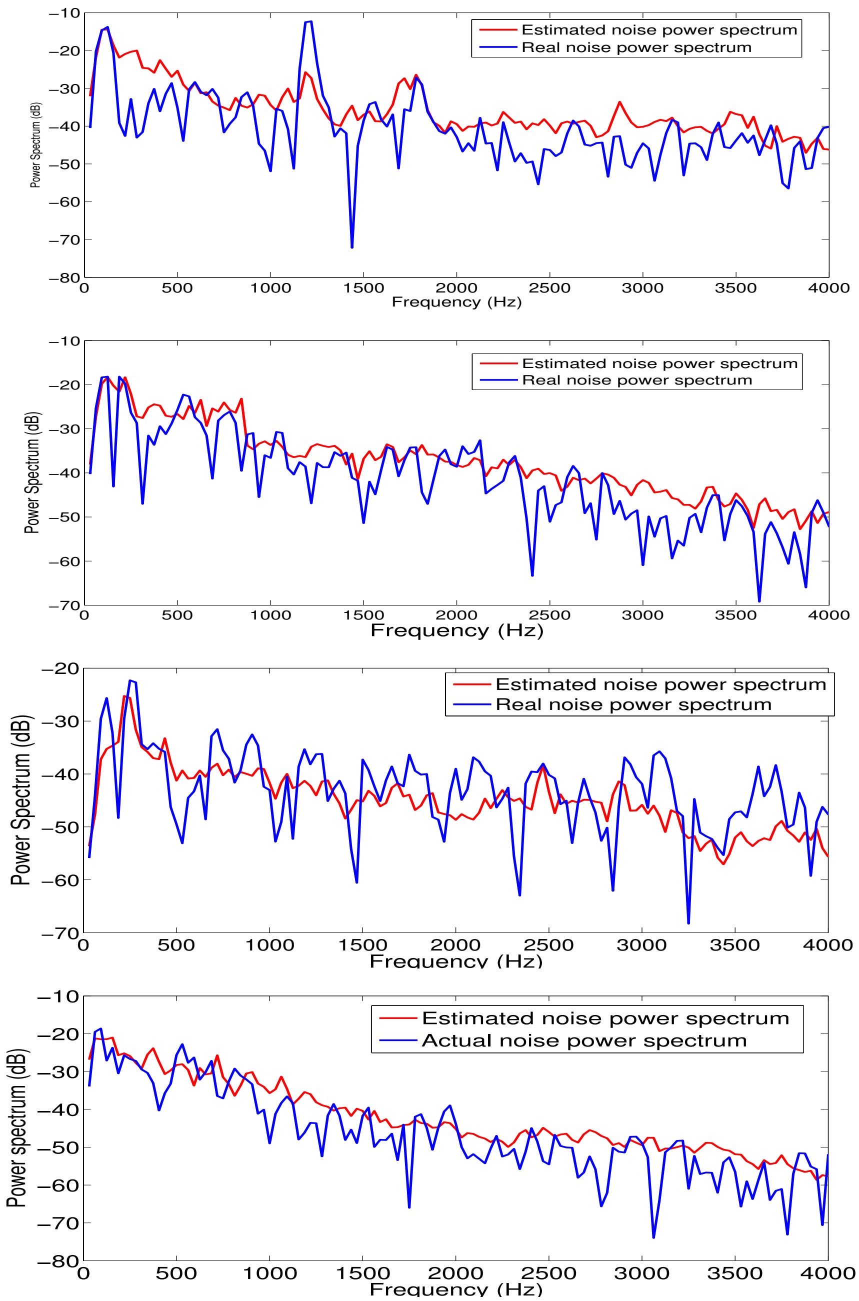


Fig. 4.3 Plot of the true and estimated noise power spectra, for female speaker at SNR=0dB. From top to bottom: train noise, car noise, street noise and airport noise

Another series of experiments have been carried out to evaluate the performance of the codebook-based noise estimation algorithm for different SNRs. Figure 4.4 compares the results of the codebook-based noise PSD estimation to the corresponding periodograms of the noisy speech.

The methodology is similar to that is Figure 4.3, except that here, we fix the noise type and vary the SNR. Specifically, we consider speech from a male speaker, contaminated by train noise at SNR=0.5 and 10dB. These results demonstrate the ability of the codebook based approach to properly estimate the noise envelope over a wide range of SNR values. Figure 4.5 present the results of a similar experiment for a different male speaker with airport noise at SNR=0dB and 10dB.

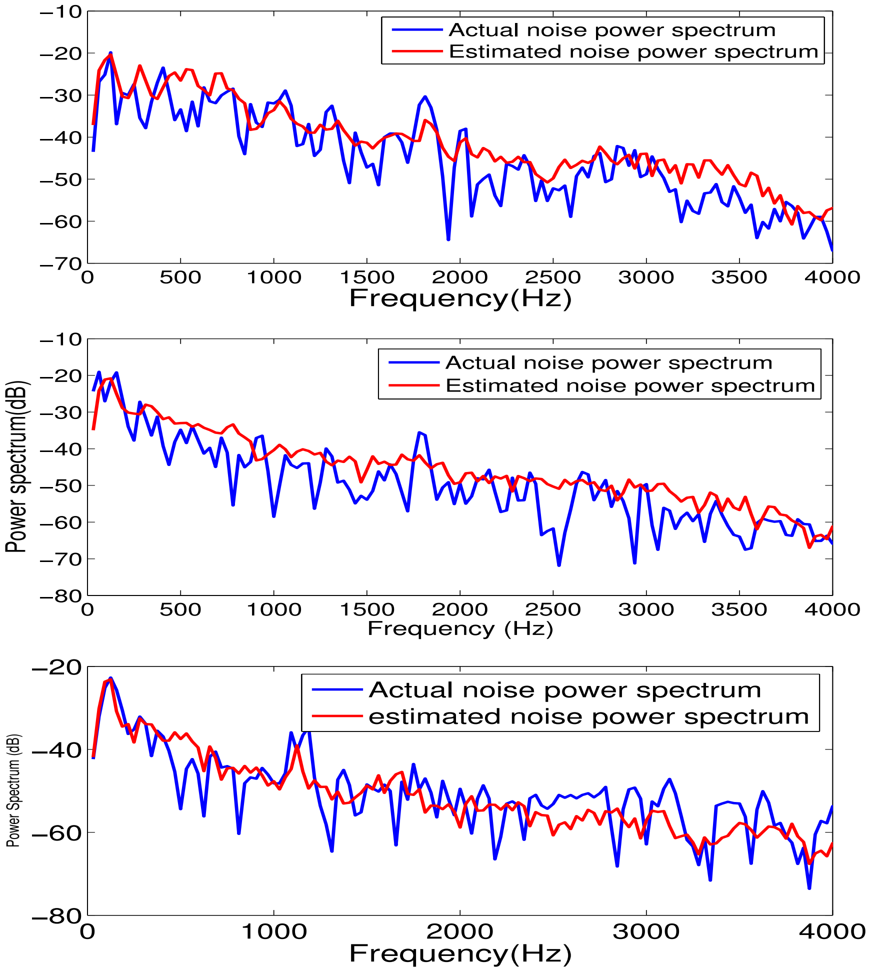
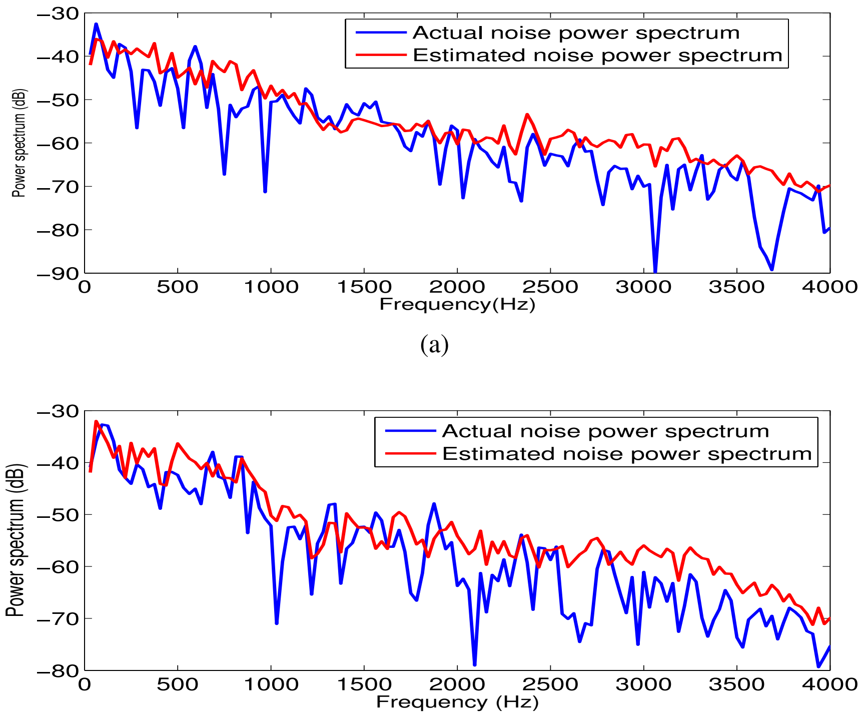


Fig. 4.4 Plot of the true and estimated noise power spectra, for a male speaker contaminated by train noise. From top to bottom: SNR=0dB, SNR=5dB, SNR=10dB



(a)

(b)

Fig. 4.5 Plot of the true and estimated noise power spectra, for a male speaker contaminated by airport noise. From top to bottom: SNR=5dB, SNR=10dB

4.2.3 Enhanced speech results

In this section, the quality of the enhanced speech estimates obtained with the proposed algorithm is examined. The time-domain signal waveforms of the noisy, true and the enhanced speech are plotted in Figure 4.7 in order to test the algorithm's efficiency. The results correspond to the speech of a male speaker, contaminated by the street noise at 5dB SNR. As it can be observed, the proposed algorithm performs relatively well in removing street noise from the noisy speech.

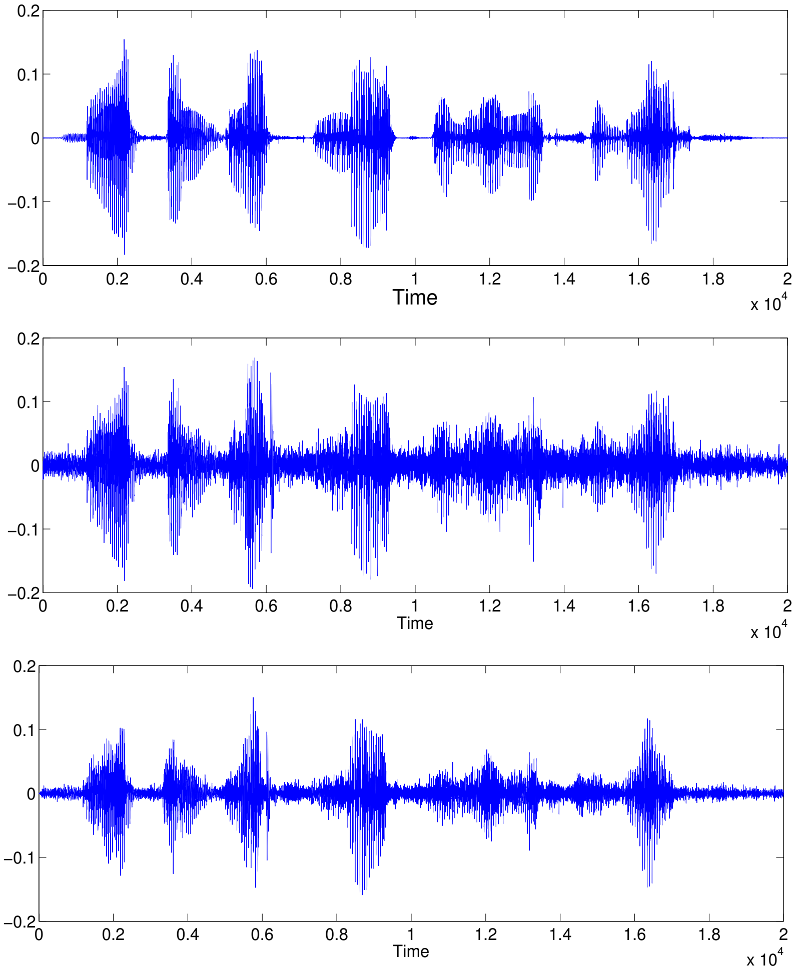


Fig. 4.6 Time domain waveforms for a male speaker and street noise at SNR=5dB. From top to bottom: clean speech, noisy speech, enhanced speech

The same experiment is repeated for a different scenario where the speech of a female speaker is contaminated by train noise at 10dB SNR.. The results are given in Figure 4.7 . It can be observed that certain amount of noise has been removed from the noisy speech signal.

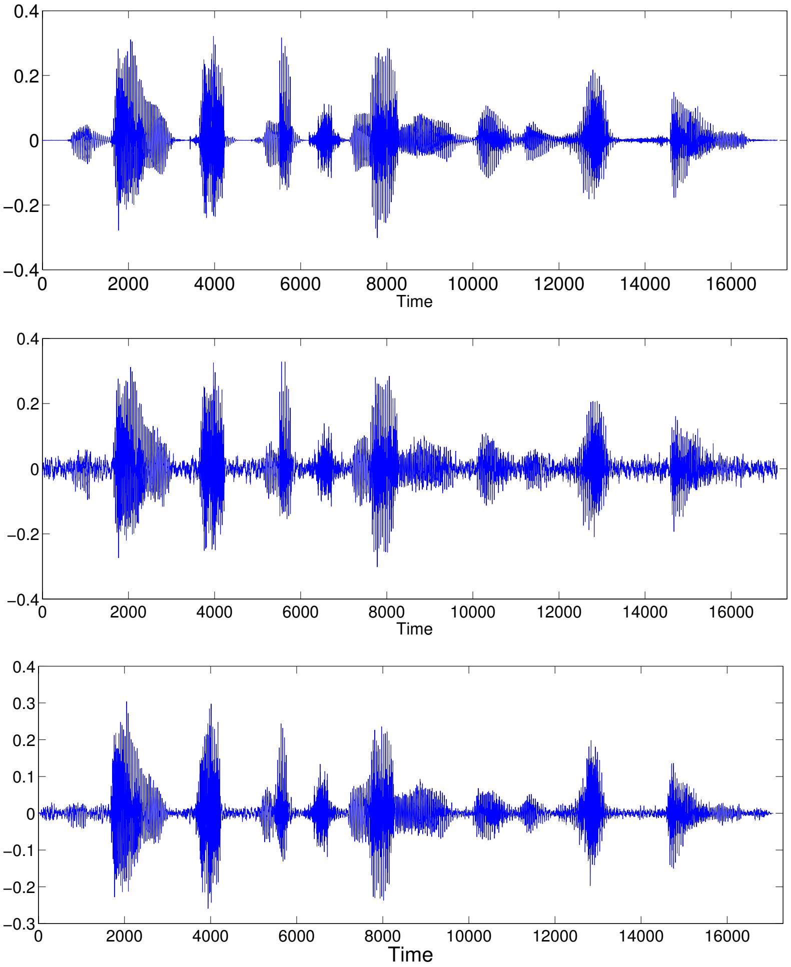


Fig. 4.7 Time domain waveforms, for a female speaker and train noise at SNR=10dB. From top to bottom: clean speech, noisy speech, enhanced speech

4.3 Objective Measure Results

In this section, the results obtained from the two speech enhancement algorithms, obtained by Wβ-SA and Wiener filter with the codebook-based MMSE noise PSD estimation approach, are compared in terms of the PESQ objective measure. At first the codebook-based method is applied in order to obtain the speech and noise PSD, and then the acquired PSDs are used in the Wβ-SA and Wiener filter methods to obtain the enhanced speech. In each table presented in this section, the PESQ objective measure is examined for the speech of a different speaker contaminated by four different types of noise at various SNRs.

In the Table 4.1 the PESQ objective measure is presented for the case where the speech of a female speaker is degraded by four different types of noise, including train, airport, car and street noise. The results are given for 3 different values of SNR, i.e. 5dB and 10dB. In tables 4.2, 4.3 and 4.4, we present similar results for the second female, the first male and the second male speakers, respectively.

Table 4.1 PESQ objective measure for enhancement of noisy speech from first female speaker

	Train noise		Airport noise		Street noise		Car noise	
	Wiener	Wβ-SA	Wiener	Wβ-SA	Wiener	Wβ-SA	Wiener	Wβ-SA
SNR=0dB	1.41	1.61	1.32	1.80	0.25	0.39	1.49	1.73
SNR=5dB	1.60	2.31	1.98	2.02	0.23	0.97	1.71	2.05
SNR=10dB	2.30	2.40	2.19	2.35	1.92	2.21	2.07	2.35

Table 4.2 PESQ objective measure for enhancement of noisy speech from first female speaker

	Train noise		Airport noise		Street noise		Car noise	
	Wiener	Wβ-SA	Wiener	Wβ-SA	Wiener	Wβ-SA	Wiener	Wβ-SA
SNR=0dB	1.78	1.94	1.78	2.10	1.45	1.66	1.59	2.05
SNR=5dB	1.94	2.42	2.52	2.62	2.01	2.32	1.92	2.43
SNR=10dB	2.26	2.68	2.36	2.58	2.47	2.50	2.32	2.59

Table 4.3 PESQ objective measure for enhancement of noisy speech from first female speaker

	Train noise		Airport noise		Street noise		Car noise	
	Wiener	Wβ-SA	Wiener	Wβ-SA	Wiener	Wβ-SA	Wiener	Wβ-SA
SNR=0dB	1.67	2.40	1.99	2.17	1.96	2.21	1.64	2.11
SNR=5dB	2.15	2.46	2.52	2.69	1.89	2.41	2.37	2.59
SNR=10dB	2.73	2.86	2.51	2.63	2.37	2.59	0.83	0.70

Table 4.4 PESQ objective measure for enhancement of noisy speech from first female speaker

	Train noise		Airport noise		Street noise		Car noise	
	Wiener	Wβ-SA	Wiener	Wβ-SA	Wiener	Wβ-SA	Wiener	Wβ-SA
SNR=0dB	1.64	2.12	1.74	1.98	1.38	1.63	1.55	1.81
SNR=5dB	2.07	2.15	1.93	2.21	1.83	2.04	2.06	2.22
SNR=10dB	1.99	2.46	2.37	2.52	2.32	2.39	2.26	2.40

According to the results, the proposed algorithm is superior to the Wiener filter in all cases expect for one (Table 4.3, car noise, 10dB), in terms of the PESQ measure. On average, PESQ measure is improved by 0.23 when the proposed method is used compared to the combination of codebook-based method and the Wiener filter.

4.4 Subjective Measure Results

The enhanced speech obtained using the described codebook-based method incorporated into the Wβ-SA speech enhancement technique was compared to the enhanced speech obtained using the codebook-based method combined with Wiener filter speech enhancement system. The results were evaluated by 10 students in Telecommunications and Signal Processing laboratory at McGill university. The methods were evaluated by pairwise comparisons to each of the noisy utterances. Almost in all cases, the students preferred the sound files obtained with the proposed method over the results obtained by the combination of codebook-based and Wiener filter.

Chapter 5

Summary and Conclusion

5.1 Summary and Conclusion

There exist several commercial systems where the removal of background additive noise from a speech signal is desirable. These include sound recording, cell phones, hands-free communications, teleconferencing, hearing aids, and human-machine interfaces such as an automatic speech recognition system [3]. Over the years, many speech enhancement approaches have been proposed to remove additive noise including spectral subtraction [27], [28], Wiener filtering [29] and Bayesian approaches [30],[31],[32].

In Chapter 1, an introduction to the speech enhancement problem and the different methods available for its solution were presented. Among all these different methods, Bayesian approaches are found to be superior to others in terms of the overall quality of the enhanced speech, the amount of speech distortion introduced by the processing and the background noise reduction. In general, in the Bayesian estimation approach for single-channel speech enhancement, an estimate of the clean speech in the frequency domain is derived by minimizing the expectation of a cost function which represents the error between the estimated and the real speech. This leads to computing a gain function which is then applied to the spectrum of the noisy speech in order to derive and estimate of the spectrum of the enhanced speech. The MMSE estimator is one of the most well-known Bayesian estimators, in which the cost function is the squared error between the estimated and actual clean speech STSA [30]. Subsequently other Bayesian estimators were developed by generalizing the MMSE STSA method, including the log-MMSE [31], WE [32], β-SA [38] and Wβ-SA [39] methods.

In Chapter 2 we began by reviewing two noise PSD estimation methods, namely the minimum statistics and IMCRA method. This was followed by a detailed discussion of the Bayesian speech enhancement algorithms. As explained, the WE estimator [32] incorporates a weighting factor while the β-SA estimator [38] incorporates a power law in the definition of their cost function. In the Wβ-SA estimator [39], the power law of the β-SA estimator and the weighting factor of the WE estimator are combined to build the cost function. The parameters (i.e. β and α) in the gain function of the Wβ-SA estimator are chosen according to characteristics of the human auditory system, namely, the compressive nonlinearities of the cochlea, the perceived loudness and the ears masking properties. Compared to other Bayesian estimators, Wβ-SA is known to achieve better enhancement performance [2].

In Chapter 3 two well-known methods which use the trained codebooks of speech and noise as *a priori* knowledge of these signals were discussed. In [42], a search is done through the codebooks in order to find the excitation variances of speech and noise that maximize the likelihood function. Afterwards, the computed excitation variances along with the LP coefficients stored in each pair of speech and noise codevectors are applied to model the speech and noise power spectrum. In [44] a similar approach is followed but instead of maximizing the log-likelihood measure, a Bayesian MMSE approach based on the Itakura-Saito measure is applied. in this method, a search is performed through the speech and noise codebooks in order to find the excitation variances which minimize the Itakura-Saito measure. Afterwards, the PDF of the observed noisy speech is modeled using the ML estimates of speech and noise, and then this knowledge of observed data PDF is applied in a Bayesian MMSE approach, in which the MMSE estimates of the speech and noise LP coefficients along with their excitation variances are derived.

In this thesis, the Bayesian MMSE estimator of the speech and noise statistics is incorporated in the Wβ-SA speech enhancement method. The knowledge of the speech and noise statistics obtained by the Bayesian scheme is applied to calculate the Wβ-SA gain function in order to derive an estimate of the enhanced speech. The incorporation of the Bayesian MMSE estimator into the Wβ-SA speech enhancement method is explained in detail in Chapter 3.

In Chapter 4, we investigated the performance of the newly proposed speech enhancement method developed in Chapter 3 by presenting the results of selected experiments, including both objective and subjective performance evaluations. In all of the experiments, we compared the new method, which combines the Wβ-SA speech enhancement with the codebook-based approach for estimating the noise and speech statistics, with a similar but alternative scheme where the Wβ-SA is replaced by the Wiener filter. When compared to combination of codebook-based method with Wiener filter, the proposed speech enhancement approach gave rise to a notable improvement in terms of the quality of the processed noisy speech represented by the PESQ objective measure. Informal listening tests were also performed, where almost in all cases the listeners preferred the sound files obtained with the proposed method over the results obtained by the combination of codebook-based and Wiener filter.

5.2 Future Work

In the process of implementing the proposed method, we faced some difficulties in estimating the noise power spectrum. The PSD estimates tend to be underestimated at some point, and we had to apply a bias factor for the purpose of compensating this analogy. In future, we will try to make some modifications in order to resolve this problem.

In [43], Srinivasan et al. proposed an alternative approach in order to train the noise codebook. They proposed a classified noise codebook scheme, where multiple small noise codebooks are trained for a particular noise type. They asserted that using this method leads to lower the complexity while increasing the accuracy of the estimates of the speech and noise excitation variances. In the future work we will be focusing on using this alternative method in the process of training the noise codebook.

these authors also developed a memory based MMSE estimator in [44], where the PSD estimates of the previous method affects the estimates of the current method. They concluded that the memory based MMSE estimator can significantly reduced both the mean and the variance of the squared error, compared to the MMSE estimator without memory. In our future work, we could try to develop the memory based MMSE estimator with the purpose of estimating the speech and noise statistics, and then combine with the Wβ-SA method.

References

[1] D. O'Shaugnessy, *Speech Communications, Humans and Machine*. Addison-Wesley, 1987.

[2] E. Plourde, "Bayesian Short-time Spectral Amplitude Estimators for Single-Channel Speech Enhancement," Ph.D. dissertation, McGill University, Montreal, Canada, 2009.

[3] J. Benesty, *Noise Reduction in Speech Processing*. Springer, 2009.

[4] E. J.Benesty, J.Chen, *Speech Enhancement in STFT Domain*. Springer, 2012.

[5] G. Gannamaneni, "Acoustic Echo Cancellation Inside a Conference Room Using Adaptive Algorithms," Master's thesis, Blekinge Institute of Technology, Karlskrona, Sweden, 2012.

[6] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[7] A. Sangwan, W.-P. Zhu, and M. O. Ahmad, "Improved voice activity detection via contextual information and noise suppression," in *IEEE Int. Symp Circuits Systems (ISCAS)*, May 2005, pp. 868–871.

[8] C.-C. Hsu, T.-E. Lin, J.-H. Chen, and T.-S. Chi, "Voice activity detection based on frequency modulation of harmonics," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2013, pp. 6679–6683.

[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul 2001.

[10] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan 2002.

[11] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Signal Process. Lett.*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[12] S. Rangachari, "Noise Estimation for Highly Non-stationary Environments," Master's thesis, The University of Texas at Dallas, Dallas, U.S.A, 2004.

[13] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *ELSEVIER, Speech communication*, vol. 48, no. 2, pp. 220–231, February 2006.

[14] Y. J.Benesty, *A Perspective on Single-channel Frequency-domain Speech Enhancement*. Springer, 2011.

[15] Y. J.Benesty, J.Chen, *Microphone Array Signal Processing*. Springer, 2008.

[16] K. Paliwal and A. Basu, “A speech enhancement method based on kalman filtering,” in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 12, Apr 1987, pp. 177–180.

[17] S. Gannot, D. Burshtein, and E. Weinstein, “Iterative and sequential kalman filter-based speech enhancement algorithms,” *IEEE Trans. Speech, Audio Process.*, vol. 6, no. 4, pp. 373–385, Jul 1998.

[18] N. Ma, M. Bouchard, and R. A. Goubran, “Speech enhancement using a masking threshold constrained kalman filter and its heuristic implementations,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 19–32, Jan 2006.

[19] J. R.Hendriks, Timo-Gerkmann, *DFT-domain Single-Microphone Noise Reduction Speech Enhancement*. Morgan Claypool, 2013.

[20] A. Rezayee and S. Gazor, “An adaptive klt approach for speech enhancement,” *IEEE Trans. Speech, Audio Process.*, vol. 9, no. 2, pp. 87–95, Feb 2001.

[21] Y. Nagata, K. Mitsubori, T. Kagi, T. Fujioka, and M. Abe, “Fast implementation of klt-based speech enhancement using vector quantization,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2086–2097, Oct 2006.

[22] F. Jabloun and B. Champagne, “Incorporating the human hearing properties in the signal subspace approach for speech enhancement,” *IEEE Trans. Audio, Speech Process.*, vol. 11, no. 6, pp. 700–708, Nov 2003.

[23] S. Ou and X. Zhao, “Speech enhancement using inter-frame dependence in dct domain,” in *Int. Conf. Signal Process.*, vol. 1, 2006.

[24] I. Y. Soon, S. N. Koh, and C. K. Yeo, “Noisy speech enhancement using discrete cosine transform,” *ELSEVIER, Speech communication*, vol. 24, no. 3, pp. 249–257, March 1998.

[25] Z. Fenghua, Y. Le, W. Jian, and S. Qiang, “Speech signal enhancement through wavelet domain mmse filtering,” in *Int. Conf. Computer, Mechatronics, Control and Electronic Engineering (CMCE)*, vol. 5, Aug 2010, pp. 118–121.

[26] H. Sheikhzadeh and H. R. Abutalebi, "An improved wavelet-based speech enhancement system," in *INTERSPEECH*, 2001, pp. 1855–1858.

[27] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[28] N. Virag, "Speech enhancement based on masking properties of the auditory system," in *IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, vol. 1, 1995, pp. 796–799 vol.1.

[29] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, 1998, ch. 6.

[30] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[31] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[32] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 5, pp. 857–869, Aug 2005.

[33] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *ELSEVIER, Speech communication*, vol. 49, no. 7, pp. 588–601, Jul-Aug 2007.

[34] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoustics, Speech , Signal Process.*, vol. 30, no. 4, pp. 679–681, Aug 1982.

[35] B. J. Shannon and K. K. Paliwal, "Role of phase estimation in speech enhancement." in *Int. Conf. Spoken Language Process. (ICSLP)*, 2006, pp. 1423–1426.

[36] P. J. Wolfe and S. J. Godsill, "Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement," in *IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 2, Jun 2000, pp. II821–II824.

[37] ——, "A perceptually balanced loss function for short-time spectral amplitude estimation," in *IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, vol. 5, Apr 2003, pp. 425–428.

[38] C. H. You, S. N. Koh, and S. Rahardja, "β-order mmse spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 4, pp. 475–486, Jul 2005.

[39] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1614–1623, Nov 2008.

[40] S. Srinivasan, "knowledge-Based Speech Enhancement," Ph.D. dissertation, KTH - Royal Institute of Technology, Stockholm, Sweden, 2005.

[41] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech, Audio Process.*, vol. 6, no. 5, pp. 445–455, Sept 1998.

[42] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise ar-models for enhanced speech coding," in *IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, vol. 1, 2001, pp. 669–672 vol.1.

[43] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 163–176, Jan 2006.

[44] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 2, pp. 441–452, Feb 2007.

[45] S. P. A. Papoulis, *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Science, 2001.

[46] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[47] I. M. R. I. S. Gradshteyn, *Table of Integrals, Series, and Products*. Academic Press, 2000.

[48] T. Islam, "Interpolation of Linear Prediction Coefficients for Speech Coding," Master's thesis, McGill University, Montreal, Canada, 2000.

[49] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, Jan 1980.

[50] V. S. Rungtai, "Parallel Vector Quantization Codebook Generation," Master's thesis, Utah state university, Logan, Utah, U.S., 1991.

[51] U. Grenander and G. Szegő, *Toeplitz Forms and Their Applications*. Univ of California Press, 1958.

[52] B. H. T. Kailath, A.H. Sayd, *Linear Estimation*. Prentice-Hall, 2000.

[53] T. Rosenkranz and H. Puder, “Improving robustness of codebook-based noise estimation approaches with delta codebooks,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1177–1188, May 2012.

[54] F. Itakura and S. Saito, “A statistical method for estimation of speech spectral density and formant frequencies,” *Electronics and Communications in Japan*, vol. 53, pp. 36–43, 1970.

[55] O. Cappé, “Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor,” *IEEE Trans. Audio, Speech, Process.*, vol. 2, no. 2, pp. 345–349, 1994.

[56] P. Kabal, *TSP Speech Database*, Telecommunications and Signal Processing Laboratory of Mcgill University, 2002, http://www-mmsp.ece.mcgill.ca/Documents/Data/index.html.

[57] D. P. H. Hirsch, http://aurora.hsnr.de.

[58] Recommendation P.56, Objective Measurement of active Speech Level, ITU-T, Geneva, Mar. 1993.

[59] Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, ITU-T, 2001.